BJPS

Brazilian Journal of
Pharmaceutical Sciences

http://dx.doi.org/10.1590/s2175-97902023e23146

Article

Advancements in artificial intelligence and machine learning in revolutionising biomarker discovery

Gokuldas (Vedant) Sarvesh Raikar[1], Amisha Sarvesh Raikar[2*], Sandesh Narayan Somnache[2]

[1]Department of Computer science (Artificial Intelligence), Manipal Institute of Technology, Yelahanka, Bengaluru, Karnataka-India, [2]Department of Pharmaceutics, PES Rajaram and Tarabai Bandekar College of Pharmacy, Farmagudi-Ponda, Goa-India

The article explores the significance of biomarkers in clinical research and the advantages of utilizing artificial intelligence (AI) and machine learning (ML) in the discovery process. Biomarkers provide a more comprehensive understanding of disease progression and response to therapy compared to traditional indicators. AI and ML offer a new approach to biomarker discovery, leveraging large amounts of data to identify patterns and optimize existing biomarkers. Additionally, the article touches on the emergence of digital biomarkers, which use technology to assess an individual's physiological and behavioural states, and the importance of properly processing omics and multi-omics data for efficient handling by computer systems. However, the article acknowledges the challenges posed by AI/ML in the identification of biomarkers, including potential biases in the data and the need for diversity in data representation. To address these challenges, the article suggests the importance of regulation and diversity in the development of AI/ML algorithms.

Keywords: Biomarkers. Artificial Intelligence. Machine Learning. Multi-omics. Omics.

INTRODUCTION

A biomarker is a measurable biological indicator that reflects normal biological processes, disease progression, or the response to therapy. It can be a cell, molecule, protein, or physical sign that is objectively assessed (Biomarkers Definitions Working Group, 2001). Biomarkers offer advantages over traditional indicators of disease in that they can predict not just disease presence, but also its progression and changes in underlying biological processes (Chen, Sun, Shen, 2015) Clinical researchers are continually searching for new biomarkers, and have recently shifted focus to digital, non-traditional markers. Digital biomarkers often combine biological, neurological, socioeconomic, and environmental data to create an intermediary biomarker (Kyriazakos *et al*., 2021) Biomarker discovery has played a crucial role in advancing clinical research by enabling the design of personalized treatments and immunotherapy. In recent years, breakthroughs in genomic research and immunotherapy resulted in development of approved biomarker-based therapies (Subbiah, 2023). In the year 2021 more than two third drugs were approved based on genetic research (Ochoa *et al*., 2022). Presently area such as oncology and genetics are highly benefited due to approved biomarker-based therapies and other areas such as cardiology, nephrology and pulmonary.

OUTCOME DRIVEN BIOMARKERS

Outcome-driven biomarkers are biological measurements that can predict clinical outcomes in patients, such as disease progression or treatment response. These biomarkers can be used to guide clinical decision-making, help identify at-risk patients, and facilitate the development of new therapies. Figure 1. Shows classification of outcome driven biomarker

***Correspondence:** A. S. Raikar. Research Scholar. Dept. of Pharmaceutics. PES Rajaram and Tarabai Bandekar College of Pharmacy. Email: amisharaikarofficial@gmail.com. ORCID: https://orcid.org/0009-0009-1328-1589

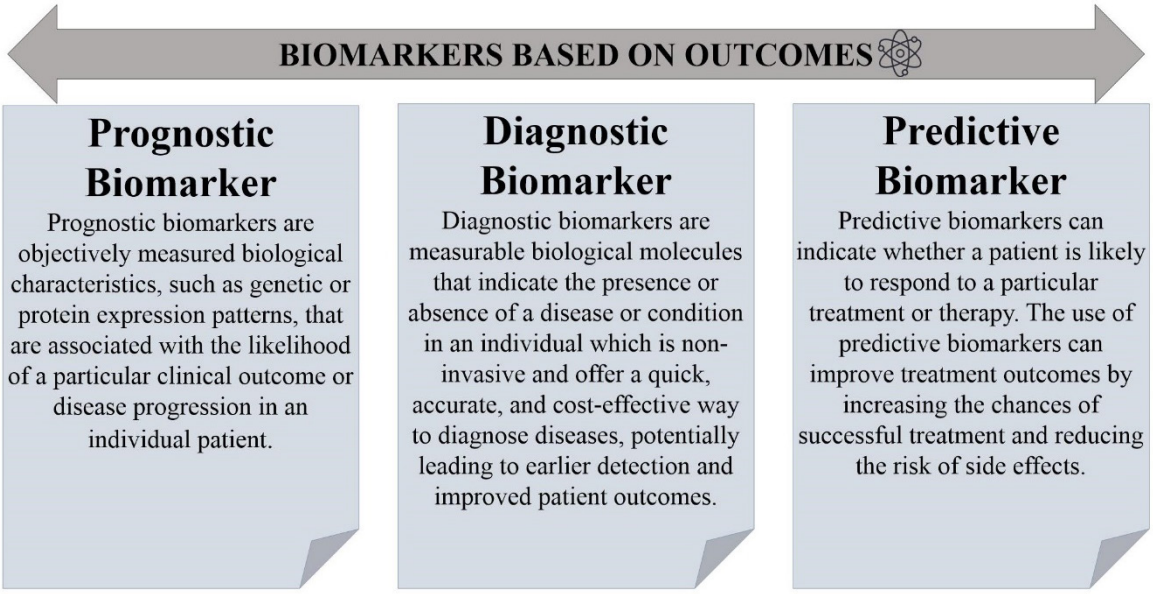

FIGURE 1 - Identifying outcome-driven biomarkers for early disease detection and therapeutic interventions.

Predictive biomarkers are a type of biological measure that can be used to anticipate patient outcomes. They assist in the selection of appropriate therapies and estimate the probability of a positive response to the treatment. For instance, mutations in the KRAS gene are often used as a predictive biomarker in colorectal cancer to determine the likelihood of response to certain therapies. Similarly, the presence of the BCR-ABL fusion gene is used as a predictive biomarker for chronic myeloid leukaemia (CML) to guide therapy decisions (de Jong *et al*., 2021). Squamous differentiation in NSCLC, CFTR Mutations, BRCA1/2 alterations and thiopurine methyltransferase genotype are the most common examples of predictive biomarkers. Squamous differentiation in NSCLC can be used to predict poor response to pemetrexed, compared to other chemotherapies such as docetaxel/cisplatin gemcitabine (Draisma *et al*., 2003). CFTR mutations helps to predict cystic fibrosis treatment efficacy (Molinski *et al*., 2018). BRCA1/2 alterations are used for prediction of platinum sensitivity in ovarian cancer (Tung, Garber, 2018). Thiopurine methyltransferase genotype used as a predictive biomarker for the treatment of 6-mercaptopurine or azathioprine, for toxicity risk evaluation (Gauba *et al*., 2006).

Diagnostic biomarkers have strong evidence of improve diagnosis accuracy. For instance, the carcinoembryonic antigen (CEA) blood test is commonly used to evaluate colorectal cancer and to determine response to treatment. Elevated levels of CEA can indicate the presence of cancer in the body (de Jong *et al*., 2021). Troponin, Procalcitonin, AFT and CEA are the most common examples of diagnostic biomarkers. Troponin can be used as a diagnostic biomarker for acute myocardial infarction (Jaffe *et al*., 2000). Procalcitonin is a biomarker for diagnosis of sepsis (Wacker *et al*., 2013). AFP serve as a diagnostic tool for liver cancer (Lok *et al*., 2010). CEA as a diagnostic marker for colorectal cancer (Saltz *et al*., 2000).

A prognostic biomarker predicts the overall outcome of a patient's illness, regardless of the therapy or treatment used. It provides insight into the natural course of the disease, allowing researchers to study and understand its underlying biological processes. In therapeutic research, prognostic biomarkers are often used in combination with predictive biomarkers to evaluate the effectiveness of various treatments. Prognostic biomarkers give information about disease outcome, independent of treatment, while predictive biomarkers predict treatment response. B-type natriuretic peptide (BNP) is

a prognostic biomarker for heart failure (Maisel *et al*., 2002), C-reactive protein (CRP) used as a prognostic biomarker for cardiovascular diseases (Ridker *et al*., 2002), Lymphocyte-to-monocyte ratio (LMR) serve as a prognostic biomarker for cancer (Zhou *et al*., 2016), Albumin-bilirubin (ALBI) grade is used as a prognostic biomarker for liver cirrhosis (Wai *et al*., 2003).

SYSTEMS BIOLOGY-DRIVEN BIOMARKERS

Another approach towards classification of biomarkers is Systems biology-driven biomarkers which are biological characteristics that are identified using systems biology approaches, which aim to understand biological systems as a whole, rather than as individual parts. These biomarkers are identified through the integration of multiple sources of data, including genomics, transcriptomics, proteomics, metabolomics, and other "omics" data, as well as clinical data and environmental factors. Examples of systems biology-driven biomarkers include gene expression signatures, metabolite profiles, and protein networks. Figure 2 classification of system biology-driven biomarkers.

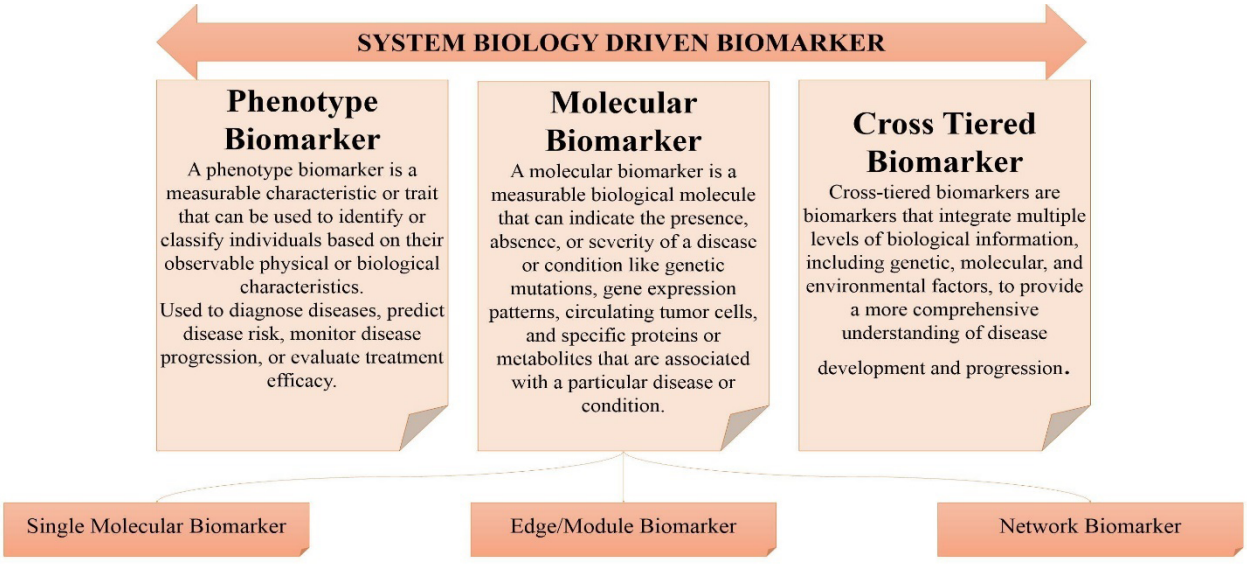

FIGURE 2 - Harnessing the power of systems biology for the discovery of novel biomarkers.

Molecular biomarkers play a crucial role in disease diagnosis and prognosis of disease by comparing the expression and concentration of individual molecules in a pathological condition. Popular methods for identifying molecular biomarkers include DESeq2 and edgeR, by analysing level of gene expression and identification of differentially expressed genes (DEGs) (Tang, Yuan, Chen, 2022).

- Single molecular Indicators, also known as node biomarkers. They are useful in diagnosing or monitoring diseases. These molecules are very sensitive and can detect a variety of conditions, such as cancer, heart disease, neurological disorders, and autoimmune diseases. Node biomarkers can also monitor disease progression and identify individuals under risk, by early diagnosis (De Fazio *et al*., 2022).
- Edge/module biomarkers are biomarkers that combines differentially expressed genes into a comprehensive framework. These biomarkers can express complex biological processes and interactions. Thus, they help in early diagnosis and effective treatments of disease (Zeng *et al*., 2016).

- Network biomarkers are involved in analysis of molecular and systems-level interactions to understand the mechanisms of health and disease. They are static (SNBs) or dynamic (DNBs) biomarker. The latter is able to track changes during disease progression and provide a comprehensive view on molecular interactions (Sonawane *et al*., 2019).

Phenotype biomarkers encompass both imaging screening and analysis of individual symptom. Imaging techniques such as CT, MRI, ECG, and EEG can identify physical anomalies, whereas analysis of individual symptom includes symptoms such as pain, fever, and bleeding. Understanding the complex pathogenic mechanisms behind changes in clinical phenotypes, such as those caused by molecular disorders, is crucial for effective medical treatments and a deeper understanding of the human body (Carr, Kraft, 2018).

Cross-Tiered biomarkers are related to environmental influences that can affect the progression of diseases. Recent research findings suggest that lifestyle habits and living environments can be key factors in the development of conditions such as cancer (Subbiah, 2023).

ARTIFICIAL INTELLIGENCE

Artificial Intelligence (AI) aims to understand and replicate intelligent behaviour by creating computer programs, seeking to uncover the underlying principles that govern both artificial and biological system. Artificial Intelligence (AI) can be classified into two different categories based on the capability and functionality. The detailed classification is presented as Figure No. 3.

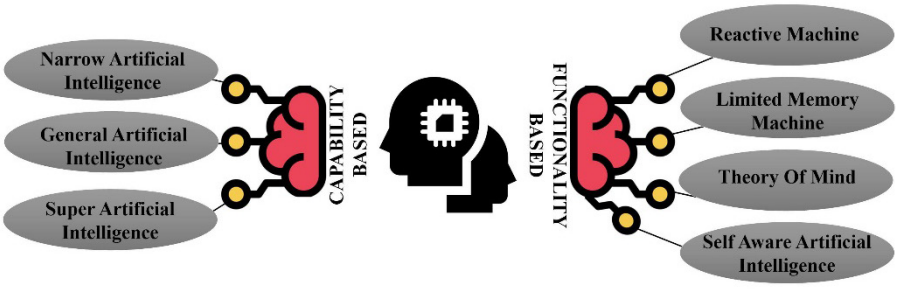

FIGURE 3 - AI classification based on capability and functionality.

Narrow AI is a weak AI, which is trained to perform specific but limited predetermined parameters. General AI is strong AI, having thinking capability. Super AI can surpass human intelligence with its capacity of thinking, reasoning, solving a puzzle, making judgments, learning, and communicating on its own.

Reactive machines respond to stimuli using pre-defined rules and lacks the ability to remember past experiences. Limited memory machines are able to learn and improve over time using data, typically by using artificial neural networks or another programming model. Deep learning, which is a subtype of machine learning, falls into this category. Theory of mind is a hypothetical stage of AI that has human-like decision-making abilities and can understand and recall emotions, as well as respond to social situations. Self-aware AI refers to machines that have an awareness of their own existence and have human-like intellectual and emotional capabilities. Theory of mind and self-aware AI are under developmental stage.

Artificial neural networks are the training models used in AI for classification and analysis of data. Some of the most common artificial neural networks are summarised in Table I.

TABLE I - Summary of Artificial Neutral Networks

\Feedforward artificial neural network	It has data flowing in one direction through layers of neurons, with the output being achieved at the end.
Recurrent neural networks	These networks use time-series data or sequences and have the ability to "remember" previous events in the sequence.
Long/Short-term Memory	This is an advanced form of RNN that can remember events from several layers ago through memory cells.
Convolutional neural networks	These networks are commonly used in image recognition and use several layers to filter different parts of an image.
Generative Adversarial Networks	These networks involve two neural networks competing with each other to improve accuracy, where one network creates examples and the other network tries to prove them true or false.

The utilization of Artificial Neural Networks (ANNs) facilitates the analysis of extensive genetic and clinical datasets, enabling the identification of interrelationships and patterns amid various variables. ANN are capable of analysing gene expression data obtained from cancer patients, with the aim of identifying new and distinctive biomarkers that are linked to specific subtypes of cancer. The identification of biomarkers through the use of artificial neural networks (ANNs) has the potential to enhance the accuracy and individualization of cancer diagnosis and treatment by furnishing precise and tailored information regarding the patient's medical status. The application of artificial neural networks (ANNs) in the sphere of biomarker discovery pertains to the discernment of biomarkers associated with prostate cancer. The application of an artificial neural network (ANN) by researchers in the examination of gene-expression data pertaining to individuals diagnosed with prostate cancer yielded a cohort of genes that demonstrated a noteworthy correlation with the progression of prostate cancer (Takeuchi *et al*., 2019). Subsequently, this particular collection of genes was subjected to validation in a distinct cohort of patients, thus indicating the promise of artificial neural networks (ANNs) in the identification of biomarkers**.** AI and machine learning have enormous potential to transform healthcare, providing faster discovery of biomarkers, improved understanding of diseases and treatments, and enabling personalized medicine. AI algorithms can analyse diverse data sources, uncover correlations, and lead to more effective treatments. Various techniques are available for analysing data, including machine learning algorithms, deep learning, natural language processing and data mining techniques. AI is also enhancing medical diagnosis, reducing diagnosis time and improving patient care [Lin *et al*., 2019)].

Natural language processing (NLP) constitutes a sub-domain of artificial intelligence (AI) focusing on the dynamic relationship and mutual communication sustained between computers and human language (Holmes *et al*., 2021). it is potentially valuable for the identification and extraction of pertinent information from diverse scientific literature, including research papers, clinical trials, and medical publications. Through the analysis of vast quantities of unstructured textual information, natural language processing methodologies may uncover associations among genetic material, proteins, and additional biological components, thus aiding in the detection of potential biomarkers.

In the study conducted in 2021 (Silverman *et al.*, 2021), highlighted the use of natural language processing (NLP) techniques for the discovery of biomarkers in the context of COVID-19 Specifically, the study compared two methods for extracting symptoms from Emergency Department (ED) admission notes, which are typically unstructured and therefore not readily available for clinical decision making. This study showcased the power of NLP techniques in biomarker discovery and has implications for the development of symptomatology-based models for other diseases.

MACHINE LEARNING MODEL

A machine learning model simulates a real-world phenomenon by utilizing training data to generate an artifact, commonly referred to as the model. This involves instructing the algorithm using a training dataset and testing the model's accuracy using a separate dataset, known as the testing data. Overfitting occurs when the model is excessively complex, while underfitting transpires when the model is inadequate in capturing patterns in the training data. Bias and variance are the two errors in machine learning models. Bias arises from incorrect assumptions of the relationship between the features and the target variable, leading to an underfit model. Variance results from the model's sensitivity to slight fluctuations in the training data, leading to overfitting. A successful machine learning model balances the trade-off between bias and variance. Supervised, unsupervised, semi-supervised, and reinforcement learning are common machine learning models for training AI. Supervised learning maps inputs to outputs using labelled training data. Unsupervised learning categorizes unlabelled data based on attributes. Semi-supervised learning utilizes a combination of labelled and unlabelled data, while reinforcement learning rewards good performance and punishes poor performance. Machine learning (ML) algorithms can analyse large and complex datasets to identify patterns and relationships between different variables, such as genomic or proteomic data, and to identify potential biomarkers. Deep learning refers to a subclass of machine learning techniques which rely on artificial neural networks for the purpose of modeling and analysing intricate data. The utilization of high-dimensional biological data, such as genomic and proteomic data, has displayed considerable potential in the identification of biomarkers through comprehensive analysis. Several deep learning techniques have been utilized in the pursuit of biomarker discovery, including:

Convolutional Neural Networks (CNNs) have demonstrated exceptional proficiency in the field of image analysis. As a type of neural network, CNNs have exhibited a heightened capability for detecting and extracting meaningful features within images. Medical imaging technologies such as MRI and CT scans have been leveraged in biomarker discovery, wherein identifying particular features linked to specific ailments is the crucial objective.

Recurrent Neural Networks (RNNs) are a class of neural networks that exhibit notable proficiency in analysing sequential data including time-series data. Biomarker discovery has utilized them to scrutinize gene expression data and discern sequential patterns that are linked to distinct maladies. Autoencoders are a type of artificial neural network utilized for unsupervised learning (Saha *et al*., 2019). These tools have been employed to detect potential patterns in biological data that may indicate the presence of particular diseases.

These profound learning models possess the capability to construct predictive models that can be utilized for the purposes of identifying and tracking the progression of various illnesses.

Supervised learning model

Supervised learning involves training a model on labelled data, where the target variables are already known. The objective of the training process is to make predictions that accurately reflect the patterns observed in new, previously unseen data, based on the information learned from the labelled training data. The model's accuracy is evaluated based on its performance on the test data, and it is refined iteratively until the desired level of accuracy is achieved. The supervised learning technique is frequently employed for diverse applications, including object recognition in images, voice recognition technology, and language processing for computers (Milali *et al*., 2020). It also has the capability to undergo training on datasets comprising of patients' health status and diverse biomarkers to anticipate the status of any disease or response to treatment based on the aforementioned biomarker data. An instance of supervised machine learning could involve the development of a model capable of predicting the probability of a patient developing Alzheimer's disease through analysis of a range of biological markers, including genetic, imaging, and cognitive biomarkers. The model will be trained using a dataset containing patient records with established

disease status and their corresponding biomarker data (Fan *et al*., 2018). This training process will facilitate the identification and analysis of discernible patterns and interrelationships that exist between various biomarkers and disease outcomes. Steps involved in supervised learning is summarised in Figure 4.

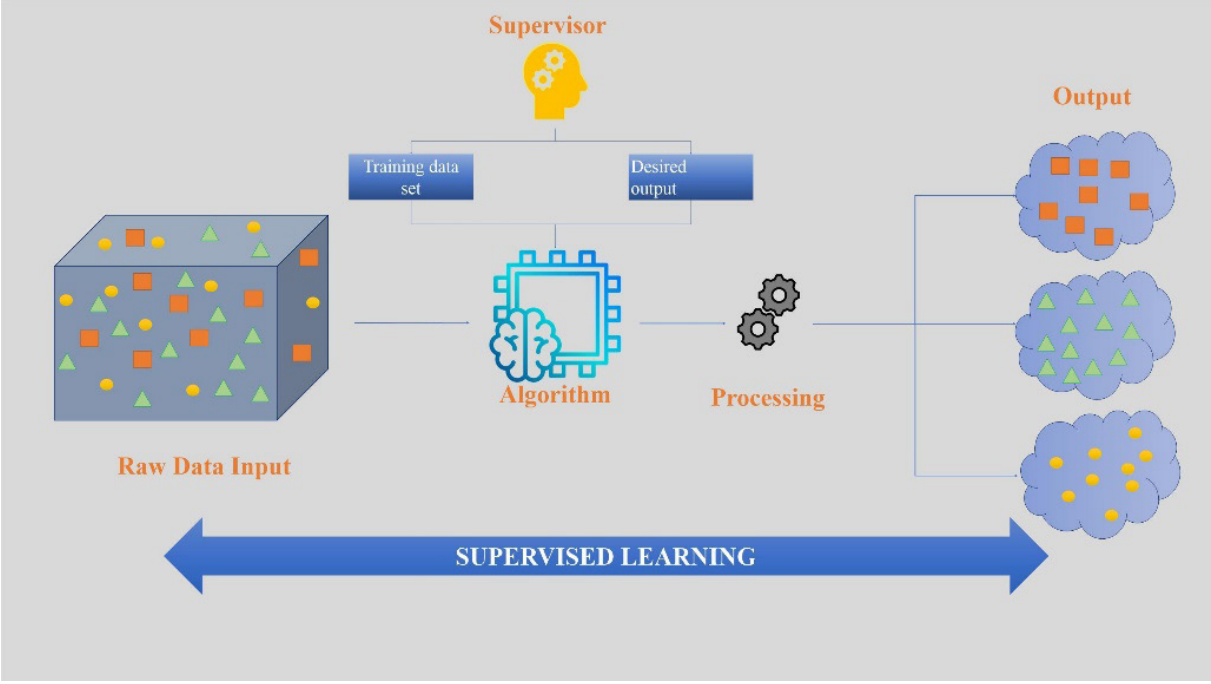

FIGURE 4 - A roadmap for supervised learning: the key stages in developing a predictive model.

The supervised learning process typically involves several steps that can vary depending on the problem and data being used. Factors such as data quality, the choice of algorithm, and the adjustment of hyperparameters can greatly impact the accuracy of the resulting model. Figure 5 shows classification of Supervised along with unsupervised learning.

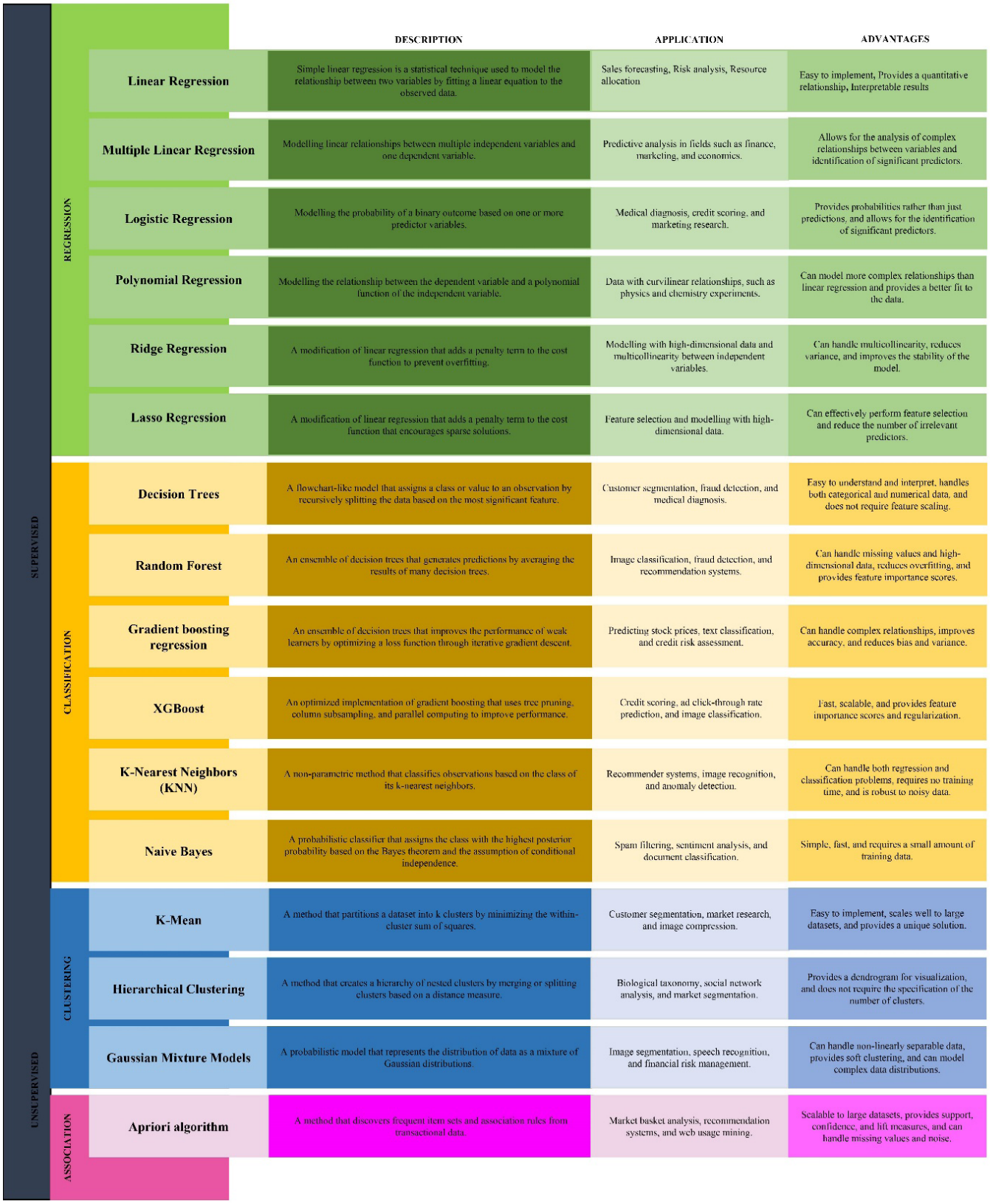

FIGURE 5 - A Guide to Supervised and Unsupervised Learning: Choosing the Right Approach for Your Data.

REGRESSION

The aim of regression is to establish a connection between the independent variables and the dependent variable, and use this connection to generate predictions for unseen data. The regression algorithm attempts to fit a line or curve that best represents the relationship between the variables. The quality of the model's predictions is typically measured using metrics such as mean absolute error, mean squared error, and R-squared (Rong, Bao-wen, 2018). It applies statistical techniques to establish a line or curve that optimally characterizes the association between two or more variables, thereby enabling researchers to make predictions regarding the values of a single variable given the underlying values of the others. In the field of biomedicine, regression models may be employed by researchers to ascertain a cohort of biomarkers that exhibit a robust correlation with the progression of a specific disease within a given patient demographic. Through the analysis of the correlation between the identified biomarkers and the clinical outcome, an effective regression model can be employed in order to ascertain the biomarkers exhibiting the highest predictive value with regards to disease progression. By utilizing these identified biomarkers, a biomarker panel may thus be developed which is capable of serving both diagnostic and prognostic objectives (Que *et al*., 2019).

CLASSIFICATION ALGORITHM

Classification Algorithm is a supervised learning technique used to categorize new data based on pre-labelled training data. It maps input to a categorical output variable with pre-defined classes, such as binary (yes/no) or multi-class (more than 2 outcomes). There are two types of learning algorithms: Lazy (stores data before testing) and Eager (develops model before testing). Examples of classification algorithms include Decision Tree, Random Forest, KNN SVM, Naive Bayes and Logistic Regression (Chowdhury, Schoen, 2020). These algorithms are generally subjected to training on a dataset that encompasses both affirmative and negative examples. The positive examples refer to individuals who exhibit the targeted condition or clinical outcome, while negative examples pertain to individuals who do not manifest the aforementioned condition or clinical outcome. The algorithm undergoes a process of learning to discern patterns present in the biomarker data and their association with positive examples. These patterns are subsequently utilized to predict the likelihood of disease manifestation or clinical outcome in unrelated patients.

UNSUPERVISED LEARNING

Machine learning that operates without labelled data is called unsupervised learning. The algorithms used in unsupervised learning work by finding patterns or structures within an unlabelled dataset. The objective of unsupervised learning is to identify relationships and connections within the data, such as grouping similar data points together into clusters. Examples of tasks accomplished through unsupervised learning include reducing the number of features in a dataset, identifying unusual or abnormal instances, and presenting data in a visual format (Amruthnath, Gupta, 2018). Popular algorithms in unsupervised learning include clustering techniques such as the K-means method, hierarchical clustering, and density-based clustering methods. These algorithms can be used to gain insights into the underlying structure of the data and make predictions based on that structure.

Previously unknown relationships between biomarkers, or to cluster patients based on their biomarker profiles can be done using unsupervised learning algorithms. Dimensionality reduction methodologies like principal component analysis, are employed to effectively decrease the number of features or variables present in a dataset, while ensuring the retention of crucial information. The utilization of aforementioned technique carries significant value in the domain of biomarker discovery as it assists researchers in pinpointing the most pertinent biomarkers for a specific affliction or clinical outcome, hence minimizing the prospect of overfitting or extraneous data in the analysis.

Clustering is a method of unsupervised learning that organizes data into groups based on similarities,

revealing patterns and structure within the data. The goal is to partition data into clusters of similar objects and is used for tasks like customer segmentation, image segmentation, and market segmentation. Some well-known clustering algorithms includes DBSCAN, Hierarchical Clustering, Gaussian Mixture Model, Affinity Propagation, Mean Shift, Spectral Clustering, Fuzzy C-Means and K-Means. This technique can be used to identify subgroups of patients with distinct biomarker profiles, which may have different disease outcomes or responses to treatment.

Association Rule Learning is an unsupervised machine learning approach to identify correlations and frequent combinations of items in large datasets. Association rule mining is a viable technique that can be employed in tandem with other methods for the purpose of biomarker discovery. It is a widely-used computational method that aims to ascertain recurring patterns or rules within abundant datasets. an approach that could be utilized is association rule mining for the identification of biomarker sets exhibiting frequent co-expression or co-regulation. The occurrence of such co-expression or co-regulation may suggest their involvement in a common biological pathway or process.

SEMI-SUPERVISED LEARNING

Semi-supervised machine learning merges aspects of both supervised and unsupervised techniques. Instead of relying solely on labelled or unlabelled data, it utilises a modest quantity of labelled data alongside of unlabelled information. This approach offers the advantages of both unsupervised and supervised learning while mitigating the difficulties of obtaining a large amount of labelled data (van Engelen, Hoos, 2020). This method enables the training of a model with fewer labelled examples, thus making the process more efficient.

Semi-supervised learning leverages the benefits of both supervised and unsupervised learning, utilizing both labelled and unlabelled data. This approach requires less labelled data compared to purely supervised learning. Compared to supervised learning and makes use of a large quantity of unlabelled data. To do this, the process of pseudo labelling is utilized. The steps in the process are given in the Figure 6.

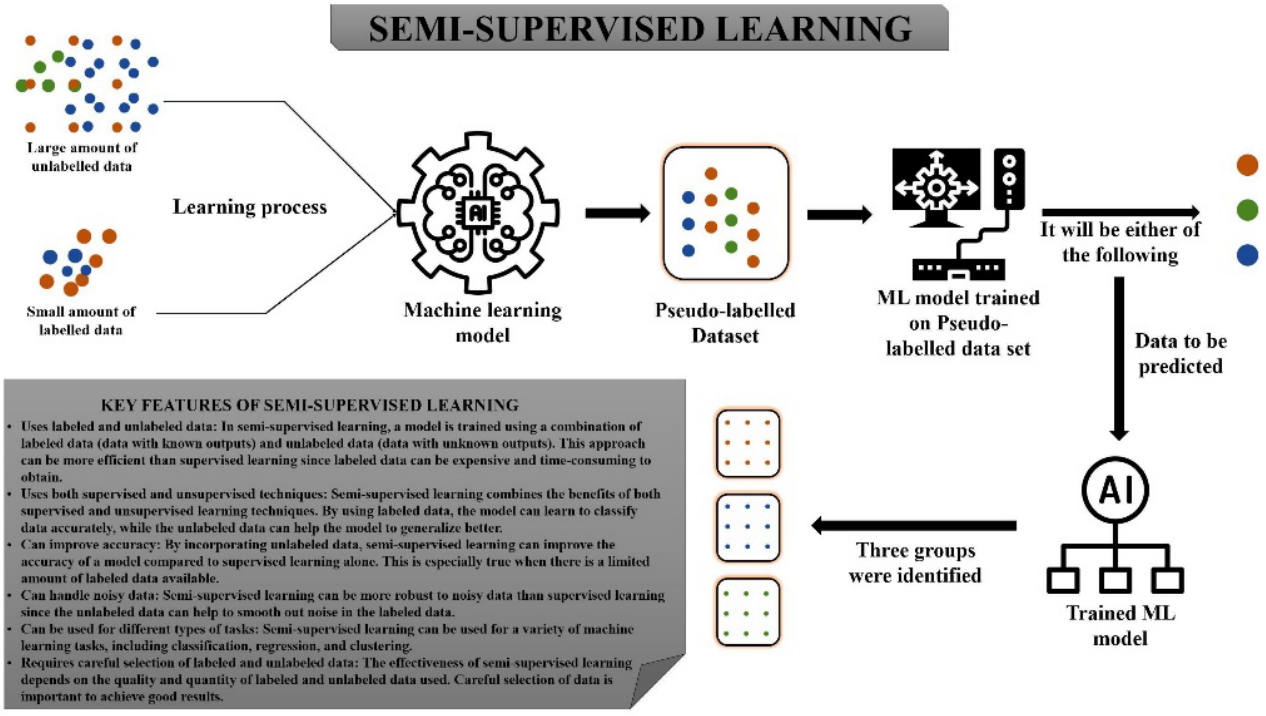

FIGURE 6 - The Power of Semi-Supervised Learning: Using Unlabelled Data to Enhance Model Accuracy.

REINFORCEMENT LEARNING

Reinforcement learning (RL) is a method of artificial intelligence which focuses on teaching a machine to make decisions by rewarding and punishing its actions. In RL, an agent interacts with its environment and learns to choose the best course of action to maximize its reward. This process is similar to how a child learns through exploration and trial-and-error. The key difference between RL and other forms of machine learning is that RL does not rely on a supervisor to provide feedback, but instead learns through its own interactions with the environment. As a result, RL is capable of discovering optimal behaviour in an unseen environment and has the potential to be highly effective in a variety of real-world applications (Hammoudeh, 2018).

Reinforcement Learning (RL) is a learning paradigm where an agent takes actions to attain a goal, and the environment provides rewards or penalties to guide the agent's decision-making process. The agent's behaviour is guided by a policy and the goal is to maximize cumulative rewards. In Reinforcement Learning, the value function is a measure of the worth of a state, and the goal is to determine the best policy that results in the highest average value of the states. RL is based on Markov Decision Processes (MDP), a mathematical model for decision-making in uncertain environments. Reinforcement Learning algorithms can be separated into two types: model-free and model-based. Model-free algorithms do not build a precise representation of the environment and operate through trial-and-error. Classes of Reinforcement learning Algorithms is given in Figure 7.

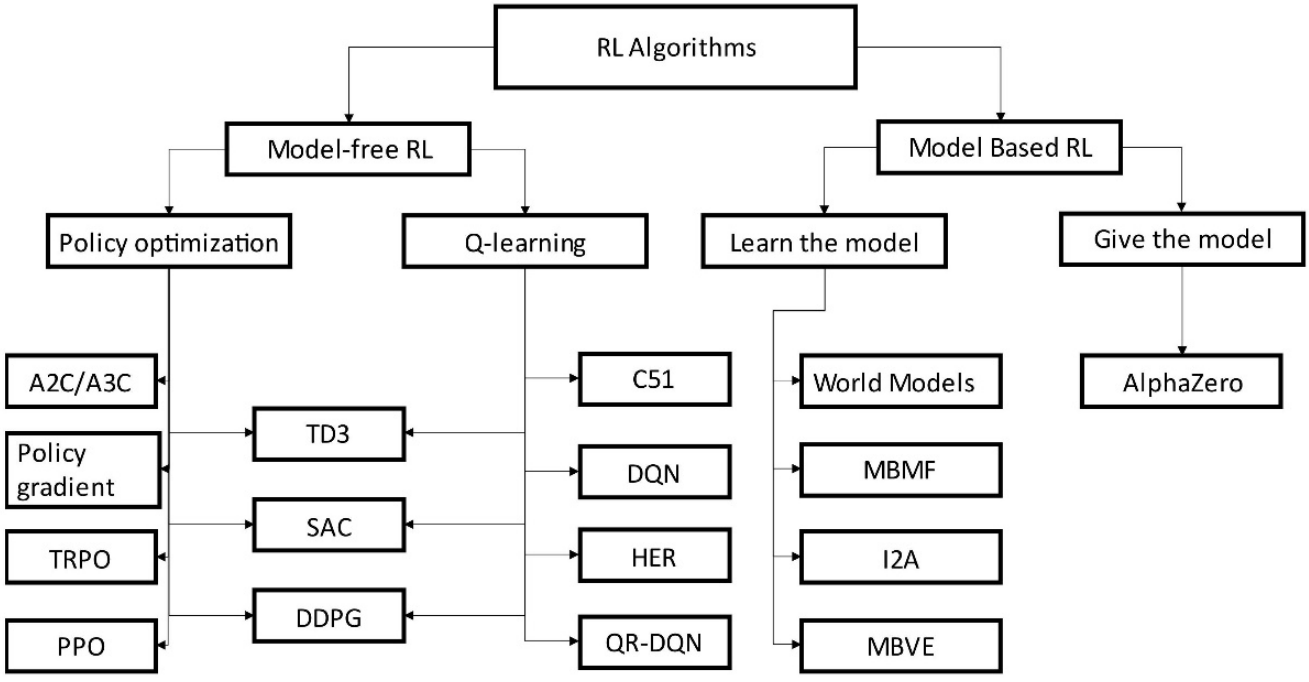

FIGURE 7 - Classifying Reinforcement Learning Algorithms: From Model-Based to Model-Free Approaches.

AI AND ML ENABLED BIOMARKERS

The digital revolution in healthcare is transforming medical research, diagnostics, and therapeutics with the widespread adoption of digital health technologies (DHTs), such as sensors and wearable devices with AI/ML capabilities. One popular DHT is wearable sensors that which provide continuous real-time monitoring and health data reporting. These sensors can track various health metrics, such as blood pressure, body temperature,

sleep quality, and cognitive decline, providing more accurate and patient-centric digital outcomes. Wearable sensors are becoming increasingly important in medical research and patient care.

Digital Biomarkers

Digital biomarkers (DB) are technological measures used to assess an individual's physiological and behavioural states, allowing for the collection of objective and quantifiable data. They utilize smart biosensors and available as wearable devices, portable instruments, implantable sensors, or ingestible tools. A well-known example of a digital biomarker is a wearable glucose monitoring device that uses a continuous glucose monitoring system to track glucose levels in real time. This type of technology provides crucial insights into a person's health, enabling healthcare providers to make more informed decisions and deliver better patient care. The advantages of digital biomarkers include: early diagnosis, remote patient care, utilizing technology, and enhancing clinical trial efficiency (Song *et al*., 2022). There are three main categories of digital biomarkers viz. Diagnostic DB, Response DB and Monitoring DB which are summarised in Table II.

TABLE II - Categories of Digital Biomarkers

Categories	Application	Example	Ref. No.
Diagnostic Digital Biomarker	To diagnose Medical Condition	Use of accelerometers and gyroscopes in smartphones and wearable devices for diagnosing and monitoring of neurological conditions such as Parkinson's disease.	(Abou *et al*., 2021)
Response Digital Biomarker	To evaluate impact of treatment	Use of smartphone-based digital biomarkers to monitor the response of patients to depression treatment. In this machine learning algorithms is used for analysis of data related to sign-symptoms, side effects, and adherence to treatment and predicting the response to the treatment.	(Mandryk *et al*., 2021)
Monitoring Digital Biomarker	To track changes in a person's health over time.	Use of wearable heart rate monitors for continuous monitoring of heart function	(Singhal, Cowie, 2020)

Digital biomarkers can be used to evaluate the effectiveness and safety of interventions. Digital biomarkers are either passive or active**.** Passive digital biomarkers, such as heart rate or oxygen saturation, are collected without patient effort. Active digital biomarkers, like tapping a button to record bladder leakage or using a smartphone camera to capture eye movements, are collected through intentional patient interaction with a device. Various biomarkers and their biosensing mechanisms are summarised in Table III.

TABLE III - Biomarkers and their biosensing mechanisms

Biomarkers	Biosensing mechanism	AI algorithms	Ref. No.
Physiological Signal	Wearable Device	Machine Learning	(De Fazio *et al*., 2022)
Glucose Levels	CGM Device	Machine Learning for Real-Time Prediction	(Galderisi *et al*., 2019)
ECG Signal	ECG Electrodes	Machine Learning for ECG Signal	(Attia *et al*., 2021)
Blood-Based Signal	Blood Testing	Machine Learning	(Putcha *et al*., 2020)
Neuroimaging	Magnetic Resonance Imaging (MRI)	Machine Learning for Neuroimaging Analysis	(Razavi, Tarokh, Alborzi, 2019)

AI-Biosensors

AI-biosensors are a combination of artificial intelligence and biosensors that have the potential to revolutionize various industries. An AI-biosensor comprises three fundamental components: data acquisition, signal transformation, and AI-based data analysis. The data collection involves using a various biosensor to monitor different types of information, which is then converted into an electrical output signal by the signal conversion system. The AI-data processing layer processes the output signal to make predictions and decisions based on the data collected by the biosensors. With their ability to provide accurate and precise measurements, AI-biosensors have applications in medical diagnostics, environmental monitoring, and more (Jin *et al*., 2020).

Wireless communication

AI-biosensors are optimally connected via wireless communication to transfer information between the biosensors and smartphone platforms or other intelligent devices. Popular wireless technologies used in AI-biosensor networks (AIBN) include Bluetooth, radio-frequency identification (RFID), near-field communication (NFC), Wi-Fi, and ZigBee (Galderisi *et al*., 2019).

Flexible electronic materials are key in integrating electronic circuits into AI-biosensors. These materials serve as a structural foundation for a range of device forms such as films, textiles, bandages, patches, and tattoos that are flexible**.** Polyimide (PI) and polyethylene terephthalate (PET) are favoured for their exceptional characteristics, comfort, high oxygen permeability, and suitability for roll-to-roll fabrication. Carbon materials are also becoming a viable option due to their light weight and high electrical conductivity. Additionally, intelligent hydrogels that alter shape and behaviours based on environmental conditions like pH, ions, temperature, and molecules are being studied. These hydrogels can even self-assemble or disassemble when interacting with the target analyte (Attia *et al*., 2021).

Smartphone-based platforms

The integration of multiple sensors and functions into smartphone-based AI biosensors has made them powerful tools for data processing, storage, sharing, and interaction with the cloud. Add-on bio-sensing modules like electrochemistry, fluorescence, and plasmon resonance along with additional hardware components enhance their functionality. Machine learning and AI enable precise detection and analysis of complex biological processes for faster and more accurate diagnoses and treatment. Past studies on smartphone-based AI biosensors are summarised in Table IV

TABLE IV - Smartphone-based AI biosensors

Type of Study	Biomarkers	Biosensing Mechanism	AI Algorithms	Ref. No.
Non-invasive measurement of glucose levels	Blood glucose levels	Reflectance spectroscopy	Artificial Neural Networks	(Zhang *et al*., 2019)
Heart rate monitoring	Heart rate	Photoplethysmography	Convolutional Neural Networks	(Coppetti *et al*., 2017)
Blood pressure monitoring	Blood pressure	Photoplethysmography	Deep Belief Networks	(Luo *et al*., 2019)
Stress level monitoring	Stress level	Photoplethysmography, Electrodermal activity	Support Vector Machines	(Gupta, Alam, Agarwal, 2020)
Sleep quality monitoring	Sleep quality	Accelerometery, Gyroscope, Magnetometer	Decision Trees	(Zhang *et al*., 2021a)

BIOMARKER IDENTIFICATION USING MACHINE LEARNING AND DEEP LEARNING

The application of Machine Learning (ML) and Deep Learning (DL) techniques is gaining increasing traction in the identification of biomarkers. Supervised and unsupervised learning algorithms are used in detecting biomarkers from diverse biological data sources. The former includes methodologies like Random Forest, SVM, Logistic Regression and Deep Learning techniques such as CNNs and RNNs (Lötsch *et al*., 2018). The latter includes PCA, ICA, and Clustering algorithms for innovative biomarker identification. Various techniques for identification of biomarkers using Machine Learning and Deep Learning are summarised in Figure No. 8

Omics data studies

Omics data study involves analysing large datasets that capture information about various biological molecules, such as genes, proteins, and metabolites. This approach can provide a comprehensive view of biological systems and help researchers understand how different molecules interact and contribute to cellular processes and disease. To analyse omics data, researchers typically use computational and statistical methods to identify patterns and relationships among different molecules and to gain insights into their functions and roles in biological systems. Figure 9. Gives an Overview of different Omics data studies

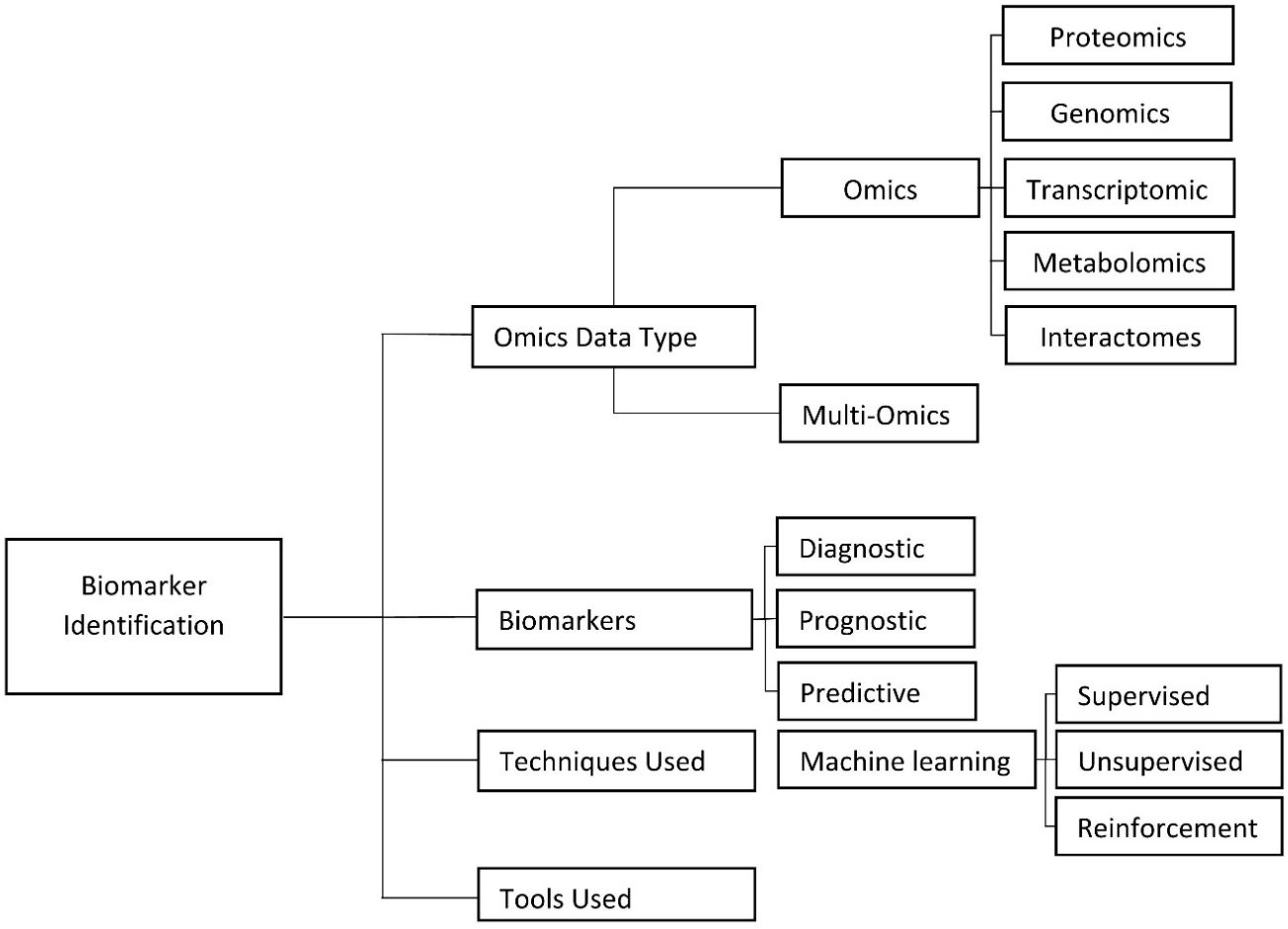

FIGURE 8 - Applying Machine Learning and Deep Learning to Biomarker Discovery: A Step-by-Step Framework.

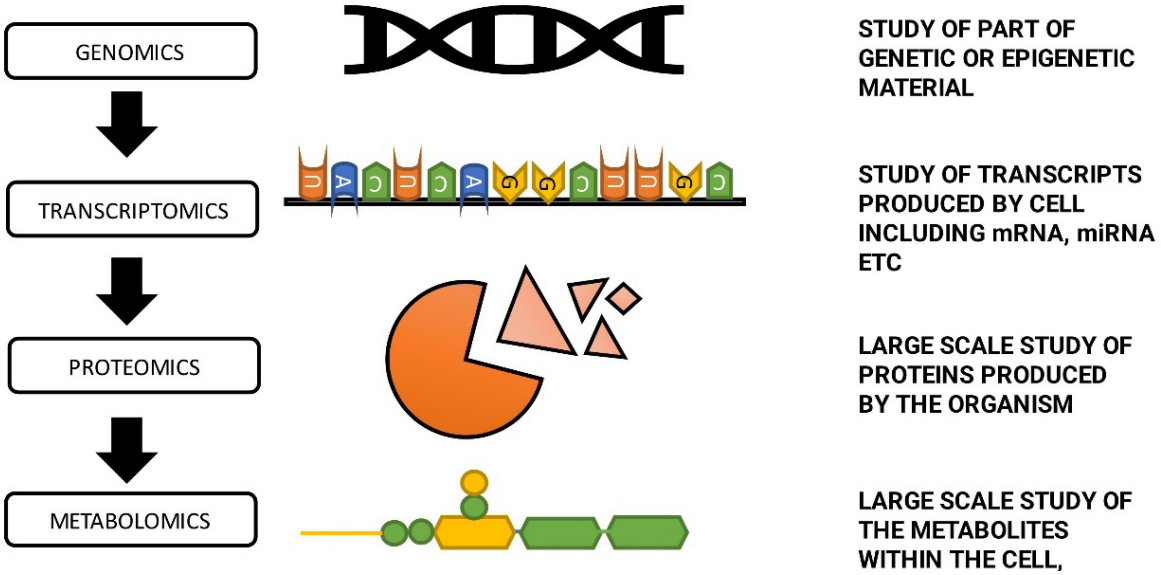

FIGURE 9 - A Comprehensive Overview of Omics Data Studies: From Genomics to Proteomics and Beyond.

Genomics: Genomics is the study of all genes within an organism, including their interactions and effects. It enables the identification of genetic variants that are linked to health, disease, and therapeutic response. Genomics provides an extensive insight into the genetic makeup of an organism and how it influences its health and well-being. The technology used for this purpose includes high-throughput sequencing, genotyping, and gene expression analysis (van Dijk *et al*., 2014). Genomic studies have changed the way we view complex diseases by identifying specific biomarkers.

Proteomics: The proteome refers to the full collection of proteins produced by a cell, tissue, or organism during a specific period. Proteomics is a technology used to investigate the expression levels, structure, and functions of proteins and to study the dynamics of their changes over time. Proteomics can be used to identify protein expression patterns in response to external stimuli, to investigate the effects of disease on the proteins expressed in a tissue, and to comprehend the complex interaction networks of proteins at the cellular and tissue level (Al-Amrani *et al*., 2021).

Transcriptomics: Transcriptomes are collections of all transcripts produced by a particular cell or tissue, including mRNA, miRNA, lncRNA, and other non-coding RNAs. RNA-seq is a powerful technique used to study transcriptomes, producing vast amounts of data that can be used to uncover new insights about gene expression, gene regulation, and other biological processes. Transcriptomes can provide valuable insights into cellular processes, gene expression patterns, and the underlying mechanisms of disease (Lowe *et al*., 2017).

Metabolome: The metabolome consists of all the metabolites, which are small-molecule groups like carbohydrates, proteinogenic amino acids, sugars, and lipid acids. Metabolomics studies can focus on different levels of metabolites and deviations or imbalances in their relative levels can indicate disease when they fall outside of normal ranges (Smirnov *et al*., 2016). There are databases, such as the Human Metabolome Database (Wishart *et al*., 2013), METLIN (Smith *et al*., 2005), and MetaboLights (Haug *et al*., 2013) that gather information on metabolites found in biological samples using techniques such as chromatography, NMR, and MS. Efforts are underway to standardize metabolomics data through initiatives like the Metabolomics Standard Initiative and the Coordination of Standards in MetabolOmics. Changes in metabolite levels can reveal an individual's genetic makeup and environmental exposures, making them useful for diagnosing conditions and understanding the molecular pathways involved.

Multi -omics data type

In order to fully comprehend the biological processes behind diverse phenotypes, a multi-omics approach is often required. Integrating multiple omics data can be a difficult task due to its large size, measurement variations, and data analysis complexity. There are various approaches to integrating omics data that have been proposed in the literature, including matrix factorization algorithms, pathway-based data integration, supervised learning methods, correlation analysis, computational frameworks, and Random Forests. Integrating other unstructured clinical phenotypic datasets, such as medical images, electronic health records, and questionnaires, also presents new challenges and requires standardization. Strategies to incorporate these data include converting them into numerical features, such as word vectors using word2vec models. The past studies carried out are summarised in Table V.

TABLE V - Past research for Biomarker identification using Machine learning and Deep learning

Study type	Disease type	Tools used	Method language	Input data used	No. of samples	Input data source	Ref.
Multiomics analysis for Alzheimer's disease biomarker prediction	Alzheimer's disease	MOFA (Multi-Omics Factor Analysis)	MOFA package in R and Python software.	metabolomics and proteomics data	120	Alzheimer's Research and therapy	(Clark *et al*., 2021)
Integrative Multiomics analysis for prostate cancer biomarker prediction	Prostate cancer	Interrogative Biology platform, Bayesian Network Inference (BNI) modules, Elastic-net coupled with bootstrapping.	R	Proteomics, metabolomics, and phospholipids	382	Journal of translation medicine	(Kiebish *et al*., 2020)
Multiomics analysis for colorectal cancer biomarker prediction	Colorectal cancer	Enrichment analyses of the co-expressed proteins	Python	proteomic and transcriptomic datasets	538	Journal of translation medicine	(Zhang *et al*., 2021a)
Integrative Multiomics analysis for breast cancer biomarker prediction	Breast cancer	GO enrichment analysis	R and python	transcriptomics, epigenomics, and proteomics data	2832	The Cancer Genome Atlas (TCGA) Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)	(Fan *et al*., 2022)
Multiomics analysis for lung cancer biomarker prediction	Lung cancer	GO enrichment analysis and KEGG analysis	Python	Genomics	189	The Cancer Genome Atlas (TCGA)	(Li *et al*., 2022)

BIOMARKER IDENTIFICATION PROGNOSTICS, DIAGNOSTICS, PREDICTIVE APPROACH USING ML

Various machine learning algorithms, namely Random Forest, Support Vector Machines, K-Nearest Neighbours, Artificial Neural Networks, and Logistic Regression, have been utilized for the purpose of identifying diagnostic, prognostic, and predictive biomarkers. Transcriptomic, proteomic, metabolomic, and epigenetic data sets derived from diverse biological samples have been employed as dynamic avenues to

uncover biomarkers potentially linked to assorted pathological manifestations such as disease onset, evolution, and therapeutic reactions. The biomarkers that have been identified possess the potential to facilitate the timely diagnosis of diseases, prognosticate the course of disease progression, and determine the most efficacious treatment alternatives. The utilization of machine learning algorithms in the discovery of biomarkers has yielded substantial enhancements in patient outcomes, with forecasts indicating its potential for transformative impact on the domain of healthcare. Studies conducted on Biomarker identification using ML and DL from multi-omics data are given in Table VI

TABLE VI - Studies conducted on Biomarker identification using ML and DL from multi-omics data

Approach	Type of biomarkers	Algorithm	Data source	Results	Ref.
Diagnostic biomarkers	Transcriptomic biomarker	Random Forest	Transcriptomics data from blood samples of patients with a particular disease and healthy controls.	Random Forest was used to predict disease status based on the expression levels of transcripts. Several transcripts were found to be associated with disease status, and these transcripts were used as potential transcriptomic biomarkers for disease diagnosis	(Huseby *et al*., 2022)
	Metabolomic biomarker	K-Nearest Neighbors	Metabolomics data from urine samples of patients with a particular disease and healthy controls.	K-Nearest Neighbours was used to predict disease status based on the levels of metabolites. Several metabolites were found to be associated with disease status, and these metabolites were used as potential metabolomic biomarkers for disease diagnosis	(Gowda *et al*., 2008)
Diagnostic biomarkers	Proteomic biomarker	Support Vector Machines	Proteomics data from serum samples of patients with a particular disease and healthy controls.	Support Vector Machines was used to predict disease status depending on the number of proteins. Several proteins were found to be associated with disease status, and these proteins were used as potential proteomic biomarkers chemotherapy resistance prediction in small cell lung cancer.	(Han *et al*., 2012)

TABLE VI - Studies conducted on Biomarker identification using ML and DL from multi-omics data

Approach	Type of biomarkers	Algorithm	Data source	Results	Ref.
Prognostic biomarkers	Transcriptomic biomarker	Artificial Neural Network	Transcriptomics data from tissue samples of patients with a particular disease and healthy controls.	Artificial Neural Network was used to predict disease progression based on the expression levels of transcripts. Several transcripts were found to be associated with breast cancer progression, and these transcripts were used as potential transcriptomic biomarkers for breast cancer prognosis.	(Chen *et al*., 2021)
	Metabolomic biomarker	Logistic Regression	Metabolomics data from blood samples of Alzheimer's patients and healthy controls.	The study found that levels of certain metabolites, such as choline and lactate, were significantly different between Alzheimer's patients and healthy controls. Logistic Regression was used to build a model that could determine Alzheimer's disease based on these metabolites. The model was able to accurately predict Alzheimer's disease with high sensitivity and specificity.	(Wang *et al*., 2021)
	Transcriptomic biomarker	Random Forest	Transcriptomics data from blood samples of patients with breast cancer and healthy controls.	The study found that levels of certain transcripts were significantly different between breast cancer patients and healthy controls. Random Forest was used to build a model that could predict breast cancer prognosis based on these transcripts. This was able to accurately determine the prognosis of breast cancer accurately	(Zare, Postovit, Githaka, 2021)

TABLE VI - Studies conducted on Biomarker identification using ML and DL from multi-omics data

Approach	Type of biomarkers	Algorithm	Data source	Results	Ref.
Predictive biomarkers	Transcriptomic biomarker	Random Forest	Transcriptomics data from biopsy samples of patients with a particular disease and healthy controls.	Random Forest was used to predict treatment response based on the expression levels of genes. Several genes were found to be associated with treatment response, and these genes were used as potential transcriptomic biomarkers for rheumatoid arthritis.	(Rychkov *et al*., 2021)
	Metabolomic biomarker	Support Vector Machines (SVM)	Metabolomics data from blood samples of patients with a particular disease and healthy controls.	SVM was used to predict treatment response based on the levels of metabolites. Several metabolites were found to be associated with treatment response, and these metabolites were used as potential metabolomic biomarkers for treatment prediction.	(Liu *et al*., 2022)
	Epigenetic biomarker	Logistic Regression	Data of DNA methylation from blood mononuclear cells of patients with a particular disease and healthy controls.	Logistic Regression was used to predict treatment response based on the levels of DNA methylation. Several regions of DNA methylation were found to be associated with treatment response, and these regions were used as potential epigenetic biomarkers for treatment prediction.	(Cappozzo *et al*., 2022)

STUDIES CONDUCTED ON BIOMARKER IDENTIFICATION USING ML AND DL FROM OMICS DATA

Studies conducted on Biomarker Identification using Machine learning and Deep learning from Omics data is listed in Table VII. The biomarkers were identified using different machine learning and statistical algorithms applied to omics data such as miRNA expression levels, DNA methylation patterns, gene expression profiling, and circulating tumour DNA (ctDNA). These biomarkers were found to have potential for diagnosing, prognosticating, and predicting disease outcomes in several types of cancer, including liver, lung, colon, breast, ovarian, and non-small cell lung cancer. Additionally, protein biomarkers were identified for Alzheimer's disease. The study highlights the importance of omics data in identifying novel biomarkers for different diseases and the potential of machine learning and statistical algorithms to analyze large-scale omics data for biomarker discovery.

TABLE VII - Studies conducted on Biomarker identification using ML and DL from omics data

Approach	Type of biomarkers	Algorithm	Data source	Results	Ref.
Diagnostic biomarkers	miRNA expression levels using RNA-seq data	Machine learning algorithms.	Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease	miRNAs were found to have diagnostic potential for several types of cancers, such as liver and lung cancer	(Pandey *et al*., 2021)
	DNA methylation patterns using genome-wide	Machine learning algorithms.	Epigenetics	DNA methylation was found to have diagnostic potential for several types of cancers, such as colon and breast cancer	(Ding, Chen, Shi, 2019)
	Multiple reaction monitoring–mass spectrometry	Machine learning algorithms	Scientific Reports	Proteins were found to have diagnostic potential for several diseases, such as Alzheimer's disease	(Kim *et al*., 2022)
Prognostic biomarkers	miRNA expression using miRNA microarray data	Machine learning algorithms	Cancer Res 2008	miRNA expression was found to have prognostic potential for Human epithelial ovarian cancer.	(Yang *et al*., 2008)
	Gene methylation using DNA methylation microarray data	Statistical algorithms	Ageing	Gene methylation was found to have prognostic potential for colorectal cancer survival	(Fu *et al*., 2020)
	Tumor infiltrate lymphocytes (TILs) using immunohistochemistry	Image analysis algorithms	Frontiers In immunology	TILs were found to have prognostic potential for cutaneous melanoma patient survival	(Maibach *et al*., 2020)
Predictive biomarkers	Gene expression profiling using RNA-seq data	Machine learning algorithms	Cancers 2021	Gene expression profiling was found to have predictive potential for non-small cell lung cancer patients	(Hijazo-Pechero *et al*., 2021)
	Allele-specific copy number analysis of tumors	Statistical algorithms	Scientific Report	Genomic CNV may be a novel prognostic biomarker for WHO grade IV glioma patient outcomes.	(Buchwald *et al*., 2020)
	Circulating tumor DNA (ctDNA) using NGS data	machine learning algorithms	Life (Basel)	ctDNA was found to have predictive potential for drug response in cancer patients.	(Lin *et al*., 2021)

TOOLS AND DATABASE FOR BIOMARKER IDENTIFICATION

Various tools and databases are available to aid in the identification and analysis of biomarkers. Biomarker Base, GSEA, IPA, Pathway Studio, MetaboAnalyst, ArrayExpress, MassIVE, TPP, GEO, and TCGA are just a few examples of these tools and databases. These resources provide researchers with access to large-scale multi-omics data, as well as a range of analysis and visualization tools for genomic, transcriptomic, proteomic, and metabolomic data. There are many other

tools available such as Illumina Platinum Genomes, Proteomics DB, Metabolomics Workbench, Human Protein Atlas, The Immune Epitope Database (IEDB), Bio Mart, STRING, and Qiime2 that can be utilized for biomarker identification and analysis. The availability of these tools and databases facilitates the identification and characterization of biomarkers and enables researchers to gain insight into the molecular mechanisms underlying biological processes and diseases. Here is a list of some tools for biomarker identification:

Biomarker Base: A multi-omics database that collects and integrates various biomarker data including genomics, transcriptomics, proteomics, and metabolomics. It provides functional analysis, pathway analysis, and data visualization tools.

GSEA: A gene set enrichment analysis tool that assesses whether a specified group of genes exhibits statistically significant similarities or differences between two biological conditions (such as phenotypes). This tool helps identify pathways and biological processes that are linked to a particular phenotype of interest (Subramanian *et al*., 2005)

Ingenuity Pathway Analysis (IPA): A web-based software platform that provides knowledge-based analysis of complex biological data, including genomic, proteomic, and metabolomic data, to help researchers understand the biological significance of their data. It offers a suite of tools for pathway analysis, gene network analysis, and more (Shao *et al*., 2020)

Pathway Studio: A proprietary software platform that provides knowledge-based analysis of biological data to help researchers understand the biological significance of their data. It offers a suite of tools for pathway analysis, gene network analysis, and more (Nikitin *et al*., 2003)

MetaboAnalyst: A comprehensive web-based platform for metabolomics data analysis and interpretation. It provides a wide range of functions including quality control, normalization, statistical analysis, pathway analysis, and more (Chen *et al*., 2022)

ArrayExpress is a publicly accessible database that stores microarray gene expression data that provides access to raw and normalized data, and the ability to perform various types of analyses, including differential gene expression analysis and functional annotation (Rustici *et al*., 2012)

MassIVE: Mass Spectrometry Interactive Virtual Environment is a free access database repository for mass spectrometry proteomics data. It provides a platform for researchers to upload, store, and share large-scale mass spectrometry data sets with the scientific community. The repository is searchable and enables users to perform data analysis, visualization, and data integration. MassIVE provides access to raw and processed data, as well as to the analysis pipelines and results (Haider, Pal, 2013).

Trans-Proteomic Pipeline (TPP): It is a software suite for the analysis of mass spectrometry-based proteomics data. It provides a complete set of tools for data processing, database search, peptide and protein identification, quantification, and functional annotation. The TPP is designed to be highly flexible and customizable, allowing researchers to choose the most appropriate algorithms and parameters for their data (Käll *et al*., 2007)

Gene Expression Omnibus (GEO): is a publicly accessible resource that offers access to microarray and RNA-sequencing data generated by researchers worldwide. GEO provides a centralized resource for the storage and analysis of large-scale gene expression data sets, enabling researchers to compare and integrate data from different studies (Edgar, Domrachev, Lash, 2002)

The Cancer Genome Atlas (TCGA): The Cancer Genome Atlas (TCGA) is a comprehensive multi-omics data repository focused on understanding the molecular basis of cancer. TCGA provides access to large-scale genomic data, including DNA sequencing, gene expression, epigenetic modifications, and microRNA expression data, for multiple cancer types (Tomczak, Czerwińska, Wiznerowicz, 2015).

Other tools that can be also utilised are Illumina Platinum Genomes, Proteomics DB, Metabolomics Workbench,

Human Protein Atlas, The Immune Epitope Database (IEDB), Bio Mart, STRING, Qiime2 etc.

CONCLUSION

In conclusion, biomarkers play a crucial role in advancing clinical research and enabling the design of personalized treatments. The integration of Artificial Intelligence, Machine Learning, and Digital Biomarkers has the potential to revolutionize the field of clinical research by providing optimized disease diagnosis and treatment strategies. AI and ML algorithms have the ability to identify patterns in large data sets, discover novel biomarkers, and improve accuracy in existing biomarkers. Digital biomarkers, leveraging the widespread adoption of digital health technologies, offer objective and quantifiable data for healthcare assessment. However, challenges in AI/ML algorithm development, such as bias due to incomplete data and the need for regulation, highlight the importance of diversity in data representation. The multidisciplinary approach to healthcare, involving collaboration between specialists, is necessary to address chronic diseases and improve patient care. With the advancements in bioinformatics, biostatistics, and multiple "omics" methods, the future of clinical research looks promising, with a potential for more targeted treatments for various conditions.

ACKNOWLEDGMENT

The authors of this comprehensive review humbly extend their deepest gratitude to the individuals that have played a seminal role in its realization. Our profound thanks are extended to the esteemed experts in the field who were so generous with their time, knowledge, and expertise, and who provided immeasurable insights and counsel. Additionally, the authors are beholden to the publisher for their support and for providing a platform that has enabled our work to reach a more extensive audience. This review serves as an emblem of the dedication and perseverance of all those who have been instrumental in guiding us on this intellectually stimulating journey.

REFERENCE

Abou L, Peters J, Wong E, Akers R, Dossou MS, Sosnoff JJ, et al. Gait and Balance Assessments using Smartphone Applications in Parkinson's Disease: A Systematic Review. J Med Syst. 2021;45(9):87.

Al-Amrani S, Al-Jabri Z, Al-Zaabi A, Alshekaili J, Al-Khabori M. Proteomics: Concepts and applications in human medicine. World J Biol Chem. 2021;12(5):57-69.

Amruthnath N, Gupta T. A research study on unsupervised machine learning algorithms for early fault detection in predictive maintenance. In: Proceedings of the 2018 5th International Conference on Industrial Engineering and Applications (ICIEA); 2018; p. 355-361.

Attia ZI, Harmon DM, Behr ER, Friedman PA. Application of artificial intelligence to the electrocardiogram. Eur Heart J. 2021;42(46):4717-4730.

Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001;69(3):89-95.

Buchwald ZS, Tian S, Rossi M, Smith GH, Switchenko J, Hauenstein JE, et al. Genomic copy number variation correlates with survival outcomes in WHO grade IV glioma. Sci Rep. 2020;10(1):7355.

Cappozzo A, McCrory C, Robinson O, Freni Sterrantino A, Sacerdote C, Krogh V, et al. A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events. Clinical Epigenetics. 2022;14(1):121.

Carr TF, Kraft M. Use of biomarkers to identify phenotypes and endotypes of severeasthma. Ann Allergy Asthma Immunol. 2018;121(4):414-420.

Chen J, Sun M, Shen B. Deciphering oncogenic drivers: from single genes to integrated pathways. Brief Bioinform. 2015;16(3):413–28.

Chen X, Chen DG, Zhao Z, Balko JM, Chen J. Artificial image objects for classification of breast cancer biomarkers with transcriptome sequencing data and convolutional neural network algorithms. Breast Cancer Res. 2021;23(1):96.

Chen Y, Li EM, Xu LY. Guide to Metabolomics Analysis: A Bioinformatics Workflow. Metabolites. 2022;12(4):357.

Chowdhury S, Schoen M. Research Paper Classification using Supervised Machine Learning Techniques. In: Proceedings of the 2020 International Conference on Innovative Trends in Computer Engineering (ITCE); 2020:1-6

Clark C, Dayon L, Masoodi M, Bowman GL, Popp J. An integrative multi-omics approach reveals new central nervous

system pathway alterations in Alzheimer's disease. Alzheimers Res Ther. 2021;13(1):71.

Coppetti T, Brauchlin A, Müggler S, Attinger TA, Templin C, Schönrath F, et al. Accuracy of smartphone apps for heart rate measurement. Eur J Prev Cardiol. 2017;24(12):1287-1293.

De Fazio R, Mattei V, Al-Naami B, De Vittorio M, Visconti P. Methodologies and Wearable Devices to Monitor Biophysical Parameters Related to Sleep Dysfunctions: An Overview. Micromachines. 2022;13(10):1335.

De Jong J, Cutcutache I, Page M, Elmoufti S, Dilley C, Fröhlich H, et al. Towards realizing the vision of precision medicine: AI-based prediction of clinical drug response. Brain. 2021;144(6):1738–1750.

Ding W, Chen G, Shi T. Integrative analysis identifies potential DNA methylation biomarkers for pan-cancer diagnosis and prognosis. Epigenetics. 2019;14(1):67-80.

Draisma G, Etzioni R, Tsodikov A, Mariotto A, Wever EM, Wernert N, et al. Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context. J Natl Cancer Inst. 2003;95(18):1375-84.

Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207-10.

Fan M, Yang AC, Fuh JL, Chou CA. Topological pattern recognition of severe Alzheimer's disease via regularized supervised learning of EEG complexity. Front Neurosci. 2018;12:685.

Fan Y, Kao C, Yang F, Wang F, Yin G, Wang Y, et al. Integrated Multi-Omics Analysis Model to Identify Biomarkers Associated with Prognosis of Breast Cancer. Front Oncol. 2022;12:899900.

Fu B, Du C, Wu Z, Li M, Zhao Y, Liu X, et al. Analysis of DNA methylation-driven genes for predicting the prognosis of patients with colorectal cancer. Aging (Albany NY). 2020;12(22):22814-22839.

Galderisi A, Zammataro L, Losiouk E, Lanzola G, Kraemer K, Facchinetti A, et al. Continuous Glucose Monitoring Linked to an Artificial Intelligence Risk Index: Early Footprints of Intraventricular Hemorrhage in Preterm Neonates. Diabetes Technol Ther. 2019;21(3):146-153.

Gauba V, Saldanha M, Vize C, Saleh GM. Thiopurine methyltransferase screening before azathioprine therapy. Br J Ophthalmol. 2006;90(7):923-924.

Gowda GA, Zhang S, Gu H, Asiago V, Shanaiah N, Raftery D. Metabolomics-based methods for early disease diagnostics. Expert Rev Mol Diagn. 2008;8(5):617-633.

Gupta R, Alam MA, Agarwal P. Modified Support Vector Machine for Detecting Stress Level Using EEG Signals. In: Versaci M, editor. Computational Intelligence and Neuroscience. 2020:8860841.

Haider S, Pal R. Integrated analysis of transcriptomic and proteomic data. Curr Genomics. 2013;14(2):91-110.

Hammoudeh A. A concise introduction to reinforcement learning. Amman, Jordan: Princess Sumaya University for Technology. 2018.

Han M, Dai J, Zhang Y, Lin Q, Jiang M, Xu X, et al. Support vector machines coupled with proteomics approaches for detecting biomarkers predicting chemotherapy resistance in small cell lung cancer. Oncology Reports. 2012;28:2233-2238.

Haug K, Salek RM, Conesa P, Hastings J, Matos P, Rijnbeek M, et al. MetaboLights: an open-access general-purpose repository for metabolomics studies and associated metadata. Nucleic Acids Res. 2013;41(D1): D781-6.

Hijazo-Pechero S, Alay A, Marín R, Vilariño N, Muñoz-Pinedo C, Villanueva A, et al. Gene Expression Profiling as a Potential Tool for Precision Oncology in Non-Small Cell Lung Cancer. Cancers. 2021;13(19):4734.

Holmes B, Chitale D, Loving J, Tran M, Subramanian V, Berry A, et al. Customizable natural language processing biomarker extraction tool. JCO Clinical Cancer Informatics. 2021; 5:833-841.

Huseby CJ, Delvaux E, Brokaw DL, Coleman PD. Blood Transcript Biomarkers Selected by Machine Learning Algorithm Classify Neurodegenerative Diseases including Alzheimer's Disease. Biomolecules. 2022;12(11):1592.

Jaffe AS, Babuin L, Clinton SK, Meijers JC, Apple FS. Troponin: the marker of the millennium in acute cardiac care. Circulation. 2000;102(10):1026-1029.

Jin X, Liu C, Xu T, Su L, Zhang X. Artificial intelligence biosensors: challenges and prospects. Biosens Bioelectron. 2020;165:112412.

Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods. 2007;4(11):923-925.

Kiebish AM, Cullen J, Mishra P, Ali A, Milliman E, Rodrigues LO, et al. multi-omic serum biomarkers for prognosis of disease progression in prostate cancer. J Transl Med. 2020;18:10.

Kim Y, Kim J, Son M, Lee J, Yeo I, Choi KY, et al. Plasma protein biomarker model for screening Alzheimer disease using multiple reaction monitoring-mass spectrometry. Sci Rep. 2022;12(1):1282.

Kyriazakos S, Pnevmatikakis A, Cesario A, Kostopoulou K, Boldrini L, Valentini V, et al. Discovering Composite Lifestyle Biomarkers with Artificial Intelligence from Clinical Studies

to Enable Smart eHealth and Digital Therapeutic Services. Frontiers Digital Health. 2021;3.

Li W, Liu B, Wang W, Sun C, Che J, Yuan X, Zhai C. Lung Cancer Stage Prediction Using Multi-Omics Data. Comput Math Methods Med. 2022;2022:2279044.

Lin C, Liu X, Zheng B, Ke R, Tzeng CM. Liquid Biopsy, ctDNA Diagnosis through NGS. Life (Basel). 2021;11(9):890.

Lin Y, Qian F, Shen L, Chen F, Chen J, Shen B. Computer-aided biomarker discovery for precision medicine: data resources, models and applications. Briefings Bioinformatics. 2019;20(3):952-975.

Liu J, Huang L, Shi X, Gu C, Xu H, Liu S. Clinical parameters and metabolomic biomarkers that predict in-hospital outcomes in patients with ST-segment elevated myocardial infarctions. Frontiers Physiol. 2022; 12:820240.

Lok AS, Sterling RK, Everhart JE, Wright EC, Hoefs JC, Di Bisceglie AM, et al. HALT-C Trial Group. Des-gamma-carboxy prothrombin and alpha-fetoprotein as biomarkers for the early detection of hepatocellular carcinoma. Gastroenterology. 2010;138(2):493-502.

Lötsch J, Lerch F, Djaldetti R, Tegder I, Ultsch A. Identification of disease-distinct complex biomarker patterns by means of unsupervised machine-learning using an interactive R toolbox (Umatrix). Big Data Analytics. 2018;3(1):1-17.

Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017;13(5):e1005457.

Luo H, Yang D, Barszczyk A, Vempala N, Wei J, Wu SJ, et al. Smartphone-Based Blood Pressure Measurement Using Transdermal Optical Imaging Technology. Circ Cardiovasc Imaging. 2019;12(8):e008857.

Maibach F, Sadozai H, Seyed Jafari SM, Hunger RE, Schenk M. Tumor-Infiltrating Lymphocytes and Their Prognostic Value in Cutaneous Melanoma. Front Immunol. 2020;11:2105.

Maisel AS, Krishnaswamy P, Nowak RM, McCord J, Hollander JE, Duc P, et al. Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. N Engl J Med. 2002;346(11):1015-1021.

Mandryk RL, Birk MV, Vedress S, Wiley K, Reid E, Berger P, et al. Remote Assessment of Depression Using Digital Biomarkers from Cognitive Tasks. Front Psychol. 2021;12:767507.

Milali MP, Kiware SS, Govella NJ, Okumu F, Bansal N, Bozdag S, et al. An autoencoder and artificial neural network-based method to estimate parity status of wild mosquitoes from near-infrared spectra. PLoS One. 2020;15(6):e0234557.

Molinski SV, Shahani VM, Subramanian AS, MacKinnon SS, Woollard G, Laforet M, et al. A Comprehensive mapping of cystic fibrosis mutations to CFTR protein identifies mutation clusters and molecular docking predicts corrector binding site. Proteins. 2018;86(8):833-843.

Nikitin A, Egorov S, Daraselia N, Mazo I. Pathway Studio - TShe analysis and navigation of molecular networks. Bioinformatics. 2003;19:2155-2157.

Ochoa D, Karim M, Ghoussaini M, Hulcoop DG, McDonagh E, Dunham I. Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. Nat Rev Drug Discov. 2022;21:551.

Pandey M, Mukhopadhyay A, Sharawat SK, Kumar S. Role of microRNAs in regulating cell proliferation, metastasis and chemoresistance and their applications as cancer biomarkers in small cell lung cancer. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer. 2021;1876(1):188552.

Putcha G, Liu TY, Ariazi E, Bertin M, Drake A, Dzamba M, et al. Blood-based detection of early-stage colorectal cancer using multiomics and machine learning. InAbstract presented at: the American Society of Clinical Oncology (ASCO) GI Symposium 2020:23-25.

Que SJ, Chen QY, Zhong Q, Liu ZY, Wang JB, Lin JX, et al. Application of preoperative artificial neural network based on blood biomarkers and clinicopathological parameters for predicting long-term survival of patients with gastric cancer. World J Gastroenterol. 2019;25(43):6451-6464.

Razavi F, Tarokh MJ, Alborzi M. An intelligent Alzheimer's disease diagnosis method using unsupervised feature learning. J Big Data. 2019;6:32.

Ridker PM, Rifai N, Rose L, Buring JE, Cook NR. Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. N Engl J Med. 2002;347(20):1557-1565.

Rong S, Bao-wen Z. The research of regression model in machine learning field. MATEC Web of Conferences. 2018;176:01033.

Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, et al. ArrayExpress update--trends in database growth and links to data analysis tools. Nucleic Acids Res. 2012;41:D987-D990.

Rychkov D, Neely J, Oskotsky T, Yu S, Perlmutter N, Nititham J, et al. Cross-Tissue Transcriptomic Analysis Leveraging Machine Learning Approaches Identifies New Biomarkers for Rheumatoid Arthritis. Front Immunol. 2021;12:638066.

Saha S, Nassisi M, Wang M, Lindenberg S, Kanagasingam Y, Sadda S, et al. Automated detection and classification of early AMD biomarkers using deep learning. Scientific Reports. 2019;9(1):10990.

Saltz LB, Cox JV, Blanke CD, Rosen LS, Hecht JR, Fehrenbacher L, et al. Irinotecan plus fluorouracil and leucovorin for metastatic colorectal cancer. N Engl J Med. 2000;343(7):905-914.

Shao Z, Wang K, Zhang S, Yuan J, Liao X, et al. Ingenuity pathway analysis of differentially expressed genes involved in signaling pathways and molecular networks in RhoE gene-edited cardiomyocytes. Int J Mol Med. 2020;46(3):1225-1238.

Silverman GM, Sahoo HS, Ingraham NE, Lupei M, Puskarich MA, Usher M et al. NLP methods for extraction of symptoms from unstructured data for use in prognostic COVID-19 analytic models. J Artificial Intelligence Res. 2021;72:429-474.

Singhal A, Cowie MR. The role of wearables in heart failure. Curr Heart Failure Rep. 2020 Aug;17(4):125-132.

Smirnov KS, Maier TV, Walker A, Heinzmann SS, Forcisi S, Martinez I, et al. Challenges of metabolomics in human gut microbiota research. Int J Med Microbiol. 2016;306(5):266-79.

Smith CA, O'Maille G, Want EJ, Chuan Q, Sunia TA, Theodore BR, et al. METLIN: A metabolite mass spectral database. Ther Drug Monit. 2005;27:747-751.

Sonawane AR, Weiss ST, Glass K, Sharma A. Network Medicine in the Age of Biomedical Big Data. Front Genet. 2019;10.

Song Y, Kang K, Kim I, Kim T-J. Pathological Digital Biomarkers: Validation and Application. Appl Sci. 2022;12(19):9823.

Subbiah V. The next generation of evidence-based medicine. Nat Med. 2023;29(1):49–58.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(27):15545-50.

Takeuchi T, Hattori-Kato M, Okuno Y, Iwai S, Mikami K. Prediction of prostate cancer by deep learning with multilayer artificial neural network. Can Urol Assoc J. 2019;13(5):E145.

Tang S, Yuan K, Chen L. Molecular biomarkers, network biomarkers, and dynamic network biomarkers for diagnosis and prediction of rare diseases. Fundam Res. 2022;2(6):894-902.

Tomczak K, Czerwińska P, Wiznerowicz M. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia, 2015;(1):68-77.

Tung NM, Garber JE. BRCA1/2 testing: therapeutic implications for breast cancer management. Br J Cancer. 2018;119(2):141–152.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C et al. Ten years of next-generation sequencing technology. Trends Genet. 2014;30:418-26.

van Engelen JE, Hoos HH. A survey on semi-supervised learning. Mach Learn. 2020;109:373–440.

Wacker C, Prkno A, Brunkhorst FM, Schlattmann P. Procalcitonin as a diagnostic marker for sepsis: a systematic review and meta-analysis. Lancet Infect Dis. 2013;13(5):426-35.

Wai CT, Greenson JK, Fontana RJ, Kalbfleisch JD, Marrero JA, Conjeevaram HS, et al. A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. Hepatology. 2003;37(1):40-6.

Wang YY, Sun YP, Luo YM, Peng DH, Li X, Yang BY, et al. Biomarkers for the Clinical Diagnosis of Alzheimer's Disease: Metabolomics Analysis of Brain Tissue and Blood. Front Pharmacol. 2021;12.

Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, et al. HMDB 3.0—The human metabolome database in 2013. Nucleic Acids Res. 2013;41(D1):D801-7.

Yang N, Kaur S, Volinia S, Greshock J, Lassus H, Hasegawa H, et al. MicroRNA microarray identifies Let-7i as a novel biomarker and therapeutic target in human epithelial ovarian cancer. Cancer Res. 2008;68(24):10307-14.

Zare A, Postovit LM, Githaka JM. Robust inflammatory breast cancer gene signature using nonparametric random forest analysis. Breast Cancer Res. 2021;23(1):92.

Zeng T, Zhang W, Yu X, Liu X, Li M, Chen L. Big-data-based edge biomarkers: study on dynamical drug sensitivity and resistance in individuals. Brief Bioinform. 2016;17(4):576-592.

Zhang C, Zeng P, Tan J, Sun S, Zhao M, Cui J, et al. Relationship of problematic smartphone use, sleep quality, and daytime fatigue among quarantined medical students during the COVID-19 pandemic. Front Psychiatry. 2021a;12:755059.

Zhang R, Liu S, Jin H, Luo Y, Zheng Z, Gao F, et al. Noninvasive Electromagnetic Wave Sensing of Glucose. Sensors. 2019;19(5):1151.

Zhang W, Lin L, Xia L, Cai W, Dai W, Zou C, et al. Multi-omics analyses of human colorectal cancer revealed three mitochondrial genes potentially associated with poor outcomes of patients. J Transl Med. 2021a;19:273.

Zhou W, Dong J, Zhang Y, Sun X, Wang H, Zhang X, et al. Lymphocyte-to-monocyte ratio as a prognostic biomarker in various types of cancer: a systematic review and meta-analysis. Oncotarget. 2016;7(6):6479-88.

Received for publication on 17th March 2023
Accepted for publication on 19th June 2023