Clinical Epigenetics

REVIEW

Open Access

# DNA methylation and machine learning: challenges and perspective toward enhanced clinical diagnostics

Erfan Aref-Eshghi[1], Arash B. Abadi[2†], Mohammad-Erfan Farhadieh[3†], Amirreza Hooshmand[4†], Fatemeh Ghasemi[5], Leila Youssefian[6], Hassan Vahidnezhad[7,8,9,10], Taylor Martin Kerrins[11], Xiaonan Zhao[11], Mahdi Akbarzadeh[12], Hakon Hakonarson[7,8,13†] and Amir Hossein Saeidian[7,11*†]

## Abstract

DNA methylation is an epigenetic modification that regulates gene expression by adding methyl groups to DNA, affecting cellular function and disease development. Machine learning, a subset of artificial intelligence, analyzes large datasets to identify patterns and make predictions. Over the past two decades, advances in bioinformatics technologies for arrays and sequencing have generated vast amounts of data, leading to the widespread adoption of machine learning methods for analyzing complex biological information for medical problems. This review explores recent advancements in DNA methylation studies that leverage emerging machine learning techniques for more precise, comprehensive, and rapid patient diagnostics based on DNA methylation markers. We present a general workflow for researchers, from clinical research questions to result interpretation and monitoring. Additionally, we showcase successful examples in diagnosing cancer, neurodevelopmental disorders, and multifactorial diseases. Some of these studies have led to the development of diagnostic platforms that have entered the global healthcare market, highlighting the promising future of this field.

**Keywords** DNA methylation, Machine learning, Epigenetics, CpG sites, Clinical application

†Arash B. Abadi, Mohammad-Erfan Farhadieh, and Amirreza Hooshmand have contributed equally to this work.

†Hakon Hakonarson and Amir Hossein Saeidian have contributed equally to this work.

*Correspondence:
Amir Hossein Saeidian
amirhossein.saeidian@bcm.edu
[1] Clinical Genetics Department, GeneDx, Gaithersburg, MD, USA
[2] Department of Medicine, Division of Clinical Immunology and Rheumatology, University of Alabama at Birmingham, Birmingham, AL, USA
[3] Department of Cell and Molecular Biology and Microbiology, Faculty of Biological Sciences and Technology, University of Isfahan, Isfahan, Iran
[4] Department of Cell and Molecular Science, Faculty of Advance Science & Technology, Tehran Medical Science, Islamic Azad University, Tehran, Iran
[5] Gastroenterology & Liver Diseases Research Centre, Research Institute for Gastroenterology & Liver Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[6] Department of Pathology, Cytogenetics Laboratory, City of Hope National Medical Center, Irwindale, CA, USA
[7] Centre for Applied Genomics (CAG), The Children's Hospital of Philadelphia, Philadelphia, PA, USA
[8] Division of Human Genetics, Children's Hospital of Philadelphia, Philadelphia, PA, USA
[9] Department of Pediatrics, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA
[10] Department of Dermatology, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA
[11] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA
[12] Cellular and Molecular Endocrine Research Center, Research Institute for Endocrine Molecular Biology, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[13] Division of Pulmonary Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA, USA

## Background

Epigenetics is the study of changes in gene function that are mitotically and/or meiotically heritable and do not entail a change in DNA sequence. Generally, epigenetics basically consists of four different but interrelated key areas: DNA methylation, histone modifications, noncoding RNAs, and chromatin accessibility [1], Epigenetics has now taken center stage for the detailed elucidation of different disease pathogeneses, mainly cancer and several other genetic disorders. Research has increasingly demonstrated that variations in epigenetic states can drive cellular diversity, even among cells with identical or nearly identical genetic sequences. This field continues to reveal how these epigenetic variations influence cell function and contribute to the complexity of biological systems and their responses to environmental cues [2].

DNA methylation is a fundamental epigenetic modification involving the addition of a methyl group to the cytosine ring within CpG dinucleotides, primarily occurring in the context of CpG islands. This process is mediated by a group of enzymes known as DNA methyltransferases (DNMTs), such as DNMT1, DNMT3a, and DNMT3b. These enzymes, often referred to as "writers," use S-adenosyl methionine (SAM) as a methyl donor to catalyze the methylation process, which plays a crucial role in gene regulation, embryonic development, and genomic imprinting [3]. Methylation is vital for processes like X-chromosome inactivation and maintaining chromosome stability. The methylation marks are removed by enzymes known as "erasers," such as the ten-eleven translocation (TET) family, including TET-1, TET-2, and TET-3. These enzymes demethylate DNA by oxidizing 5-methylcytosine (5mC) into 5-hydroxymethylcytosine (5hmC), and further into 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) [4]. This dynamic balance between methylation and demethylation is crucial for cellular differentiation and response to environmental changes. During cell division, specifically meiosis, and mitosis, methylation patterns are generally preserved, crucial for genomic stability and maintaining cellular memory across generations. DNMT1 plays a pivotal role in this maintenance by recognizing hemi methylated DNA strands during DNA replication and restoring the methylation pattern on the new strand, thus preserving the epigenetic marks through cell divisions. Understanding and detecting these methylation changes, especially 5mC, is essential for both biological research and clinical applications, as disruptions in methylation patterns are associated with various diseases [5].

Machine learning (ML) has revolutionized diagnostic medicine by enabling the analysis of complex datasets to identify patterns and make predictions. Techniques such as deep learning (DL) and neural networks can process large-scale genomic, proteomic, and clinical data, facilitating early disease detection and personalized treatment plans.

Conventional supervised methods, including support vector machines, random forests, and gradient boosting, have been employed for classification, prognosis, and feature selection across tens to hundreds of thousands of CpG sites. These approaches, which can be streamlined by AutoML, serve as the foundation for the creation of tools that are applicable to clinical settings [6]. For instance, ML algorithms have been successfully employed to predict cancer outcomes and diagnose neurological disorders with high accuracy [7, 8]. ML integration in diagnostic workflows enhances precision, reduces costs, and improves patient outcomes [9].

A notable instance is the DNA methylation-based classifier for central nervous system cancers, which standardized diagnoses across over 100 subtypes and altered the histopathologic diagnosis in approximately 12% of prospective cases, accompanied by an online portal facilitating routine pathology application [10]. Genome-wide episignature analysis in rare diseases similarly utilizes machine learning to correlate a patient's blood methylation profile with disease-specific signatures and has demonstrated clinical utility in genetics workflows [11, 12]. In liquid biopsy, targeted methylation assays combined with machine learning provide early detection of many cancers from plasma cell-free DNA, showing excellent specificity and accurate tissue-of-origin prediction that enhances organ-specific screening [13, 14]. These implementations demonstrate how ML transforms methylation patterns into useful findings in cancer and medical genetics.

Deep learning improves DNA methylation studies by directly capturing nonlinear interactions between CpGs and genomic context from data. Multilayer perceptrons and convolutional neural networks have been employed for tumor subtyping, tissue-of-origin classification, survival risk evaluation, and cell-free DNA signal identification. Recently, transformer-based foundation models have undergone pretraining on extensive methylation datasets and subsequent fine-tuning for clinical applications. MethylGPT, trained on more than 150,000 human methylomes, supports imputation and subsequent prediction with a physiologically interpretable focus on regulatory regions. CpGPT exhibits robust cross-cohort generalization and produces contextually aware CpG embeddings that transfer efficiently to age and disease-related outcomes [15, 16]. These models enhance efficiency in limited clinical populations and underscore the promise for task-agnostic, generalizable methylation learners, which will be detailed in the subsequent sections of the paper.

Agentic AI is becoming a catalyst for omics analysis by combining large language models with planners, computational tools, and memory systems to perform activities like quality control, normalization, and report drafting with human oversight. Initial examples showcase autonomous or multi-agent systems proficient at orchestrating comprehensive bioinformatics workflows and facilitating decision-making in cancer. Although these methodologies are not yet established in clinical methylation diagnostics, they signify a progression toward automated, transparent, and repeatable epigenetic reporting, dependent on attaining enough dependability and regulatory supervision [17–20].

Important limitations exist, e.g., batch effects and platform discrepancies require harmonization across arrays and sequencing. Limited, imbalanced cohorts and population bias jeopardize generalizability; hence, external validation across many sites is essential. Many DL models exhibit a deficiency in clear explanations, hence limiting confidence in regulated environments; recent advancements in interpretable overlays for brain–tumor methylation classifiers represent progress toward clinically acceptable attribution of CpG features. Currently, multicancer early detection technologies highlight high specificity, but sensitivity, especially for stage I malignancies, is progressively improving. Regulatory clearance, cost-efficiency, and incorporation into clinical protocols are current priorities of evidence development [21].

Recent studies highlight the dynamic impact of epigenetics on disease processes, with research focusing on using DNA methylation biomarkers and ML-based methods to predict disease progression and treatment responses in various syndromes and cancers [22]. This field is advancing personalized medicine and has the potential to revolutionize treatment approaches and patient care [23].

Therefore, this study investigates DNA methylation profiles, commonly termed epigenetic signature, for their practical applications in clinical diagnoses. We investigate recent methodological advancements in utilizing machine learning to assess epigenetic markers, focusing on their significance for illness diagnosis and prognosis. We examine traditional machine learning, deep learning, and recent foundational models; outline agent-assisted processes; address assessment and interpretability methodologies; and highlight deployment insights from cancer, multifactorial, and rare disease genetics. The research closes by discussing present limitations and providing prospective insights on the integration of methylation-ML systems into standard clinical practice.

## DNA methylation detection techniques

The integration of ML with clinical epigenetics is propelling precision medicine to new heights, addressing the complexities of DNA methylation profiling. A wide variety of biochemical methods are used in DNA methylation studies (see Fig. 1 and Table 1 for detailed comparisons). More advanced sequencing techniques, such as whole-genome bisulfite sequencing (WGBS), and reduced representation bisulfite sequencing (RRBS), and single-cell bisulfite sequencing (scBS-Seq), provide single-base resolution of methylation patterns but demand higher costs and computational resources [24–27]. These high-throughput sequencing methods follow a rigorous pipeline involving library preparation, alignment, quality control, methylation calling, and annotation [28]. Similarly, methods like methylated DNA immunoprecipitation (MeDIP) is an enrichment-based technique that isolates methylated DNA fragments using antibodies specific to 5-methylcytosine, followed by sequencing (MeDIP-seq) to identify methylation patterns across the genome, and MethylCap-seq adopt comparable workflows for analyzing methylation data [29]. The growing interest in single-cell methylation profiling highlights it's potential to reveal methylation heterogeneity at the cellular level, offering insights into cellular dynamics and disease mechanisms [30]. Despite these advancements, hybridization microarrays such as the Illumina Infinium HumanMethylation BeadChip array remain popular for their affordability, rapid analysis, and comprehensive genome-wide coverage. These arrays are particularly advantageous for identifying differentially methylated regions (DMRs) across predefined CpG sites, as they combine efficiency with high-resolution insights into epigenetic alterations. This versatility supports a broad spectrum of experiments, from genotyping to gene expression analysis, further cementing their role in epigenetic and clinical research [31]. In addition to these methods, enhanced linear splint adapter sequencing (ELSA-seq) has emerged as a promising approach for detecting circulating tumor DNA (ctDNA) methylation with high sensitivity and specificity. By targeting specific methylation regions with bisulfate sequencing, ELSA-seq enables precise monitoring of minimal residual disease (MRD) and cancer recurrence, making it particularly suitable for liquid biopsy applications [32].

Long-read sequencing enables the analysis of DNA fragments ranging from several kilobases up to megabases, offering key advantages over short-read NGS, such as improved detection of structural variations, higher

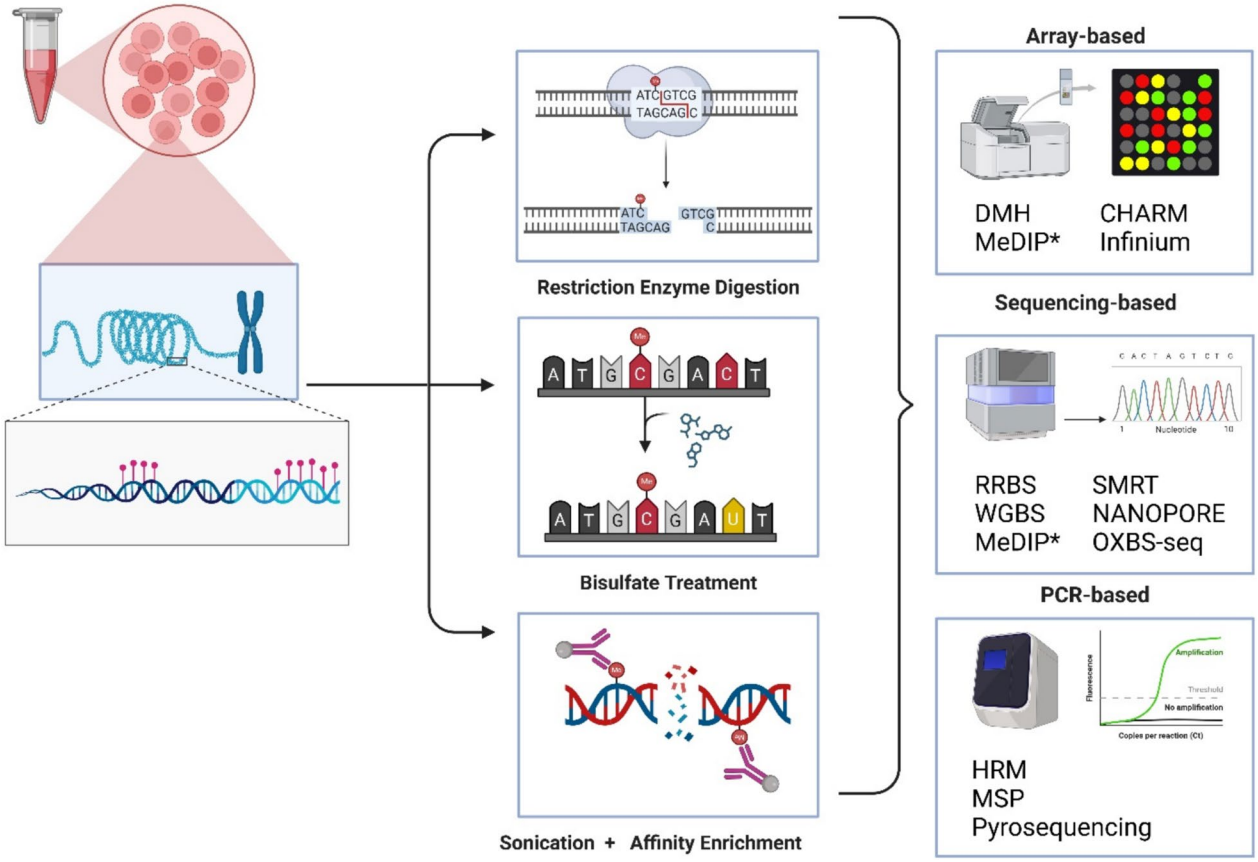


**Fig. 1** Comprehensive overview of DNA methylation analysis techniques. Workflow and various methodologies employed in genome-wide DNA methylation analysis. The process begins with the collection of DNA samples, followed by pretreatment steps such as restriction enzyme digestion, bisulfite treatment, and sonication combined with affinity enrichment. Each method targets specific methylated sites within the DNA. The techniques are categorized into array-based, sequencing-based, and PCR-based methodologies, each differing in throughput and detail of methylation data. Array-based methods include DMH and CHARM, sequencing-based cover RRBS, WGBS, and others, while PCR-based techniques encompass HRM, MSP, and pyrosequencing. *MeDIP is used in both array-based and sequencing-based methods

resolution in repetitive regions, direct identification of base modifications, and long-range haplotype phasing [33, 34]. The two primary long-read sequencing platforms are Oxford Nanopore Technologies and PacBio SMRT, each differing in throughput, error rate, and read length. Oxford Nanopore Technologies generally provides longer reads at lower cost but with slightly higher base-calling errors, although methylation detection remains highly accurate [35, 36]. Importantly, Oxford Nanopore Technologies enables real-time sequencing without the need for PCR amplification and supports direct analysis of DNA, RNA, and other biomolecules [36]. A notable application, nanoNOMe, allows simultaneous profiling of CpG methylation and chromatin accessibility, facilitating allele-specific epigenetic studies on native long DNA strands [37].

Single-cell DNA methylation profiling has emerged as a transformative approach in epigenetics, offering unprecedented resolution to investigate cellular heterogeneity, developmental processes, and disease mechanisms at the individual cell level. Techniques such as single-cell bisulfite sequencing (scBS-seq) and single-cell reduced representation bisulfite sequencing (scRRBS) enable high-resolution insights into DNA methylation heterogeneity, particularly in complex diseases like cancer and neurodegenerative diseases, where they reveal epigenetic variations driving intra-tissue heterogeneity and treatment resistance [38]. Recent advancements, such as the sci-MET method, leverage combinatorial indexing with fluorescence-activated nuclei sorting, Tn5 tagmentation, and NGS to achieve high-throughput single-cell methylome profiling [39]. Additionally, innovative non-bisulfite, enzyme-based methods like sciEM and simultaneous sc-5mC and sc-5hmC profiling via bisulfite-free chemical labeling (SIMPLE-seq) have further expanded the scope of single-cell analyses, reducing DNA degradation and improving technical feasibility [40, 41]. The scDEEP-mC method has further advanced the field by enabling efficient, high-coverage library generation, allowing for

**Table 1** Overview of DNA methylation detection techniques

| Technique | Key features | Applications | Limitations | Ref |
|---|---|---|---|---|
| High-Resolution Melting (HRM) | Rapid, cost-effective, adaptable to high-throughput | Initial screening of methylation | Does not provide site-specific methylation information | [235, 236] |
| Methylation-Specific PCR (MSP) | Uses specific primers to differentiate methylated from unmethylated DNA | Disease research, especially cancer | Requires significant DNA input; potential for false results | [237] |
| Luminometric Methylation Assay (LUMA) | Combines restriction digestion and pyrosequencing | High-throughput methylation screening | Lacks site-specific resolution | [238] |
| Methylated DNA Immunoprecipitation (MeDIP) | Enriches methylated DNA fragments via immunoprecipitation | Genome-wide methylation studies | Low resolution, depends on antibody quality | [29] |
| Pyrosequencing | Sequencing by synthesis, provides quantitative methylation data | Methylation profiling, biomarker discovery | Expensive, bisulfite treatment may degrade DNA | [239] |
| Whole-Genome Bisulfite Sequencing (WGBS) | Offers comprehensive, single-base resolution | Detailed methylation mapping across the genome | High-cost, intensive data analysis | [24] |
| Infinium Methylation 450K BeadChip | Interrogates over 450,000 CpG sites | Genome-scale methylation profiling | Cannot distinguish between methylation and hydroxymethylation<br>Less coverage than EPIC | [240] |
| Infinium MethylationEPIC arrays | Interrogates over 850,000 CpG sites<br>More sites than 450K | Genome-scale methylation profiling | Cannot distinguish between methylation and hydroxymethylation | [177] |
| Reduced Representation Bisulfite Sequencing (RRBS) | Targets CpG-rich regions, provides high coverage at lower cost | Focused methylation mapping | Limited by regions near used restriction sites | [25, 26] |
| Methylation-sensitive Restriction Enzyme Sequencing (MRE-Seq) | Uses methylation-sensitive enzymes | Methylation patterns in CpG-rich regions | Low overall genomic coverage | [241] |
| Enzyme-Linked Immunosorbent Assay (ELISA) | Quick and easy to perform | Rough estimation of global DNA methylation | High variability, low precision and sensitivity | [238] |
| Single-cell Bisulfite Sequencing (scBS-Seq) | Captures single-cell methylation heterogeneity | Studies on cellular differentiation and disease | Does not achieve full genome-wide coverage | [27] |
| Nanopore Sequencing | Uses protein nanopores for sequencing | Methylation profiling at single loci | Higher error rates, lower throughput | [242] |
| Oxidative Bisulfite Sequencing (oxBS-seq) | Distinguishes 5mC from 5hmC | Studies of DNA methylation and hydroxymethylation | Requires high coverage for quantitative sequencing | [243] |
| Methyl-HiC | Integrates Hi-C with bisulfite sequencing | Chromosomal architecture and methylation profiling | Complex and costly procedure | [244] |
| HPLC–UV | High-performance liquid chromatography-ultraviolet detection | Quantifying methylation levels | Requires professional equipment and large DNA samples | [238] |
| Liquid Chromatography Mass Spectrometry (LC–MS/MS) | Highly sensitive, can detect minor changes in methylation | Deep methylation analysis | Requires access to specialized equipment | [238] |
| AFLP, RFLP | Analyze fragments for differential methylation | Quick checks of DNA methylation | Limited analysis scope, resolution challenges | [238] |
| Differential Methylation Hybridization (DMH) | Array-based method, uses fluorescent dyes | Genome-wide methylation studies | Requires substantial DNA input, long processing times | [245] |
| Comprehensive high-throughput arrays for relative methylation (CHARM) | Uses tiling microarray to detect methylation | Gene regulation and disease development studies | Dependent on array quality, sensitive to noise | [245, 246] |
| Methylation-sensitive Single Nucleotide Primer Extension (Ms-SNuPE) | Quantitates methylation at specific CpG sites | High-throughput targeted methylation analysis | Limited to a few CpG sites per reaction | [247] |
| Matrix-assisted Laser Desorption Ionization Time-of-Flight (MALDI-TOF) | Based on bisulfite conversion | Analyzing specific gene loci or DMRs | Expensive, not suitable for genome-wide analysis | [248–250] |
| Combined Bisulfite Restriction Analysis (COBRA) | Combines bisulfite treatment and restriction digestion | Locus-specific methylation detection | Depends on the presence of restriction sites | [251, 252] |

cell-type identification, hemi-methylation profiling, and allele-resolved analysis of X-inactivation, as well as insights into DNA methylation maintenance dynamics during replication [42]. Similarly, sciMETv3 has demonstrated atlas-scale profiling capabilities, producing a dataset of over 140,000 cells from the human cortex in a single experiment, integrating chromatin accessibility and DNA methylation data from the same cells [43]. These advancements have also facilitated the creation of scalable DNA methylation atlases across 13 tissues and 40 cell types, validated using bulk and single-nucleus datasets, providing a valuable resource for biomarker discovery and methylome interpretation [44]. In neurological disorders, deep learning models like INTERACT and scMeFormer have demonstrated remarkable capabilities by using single-cell DNA methylation data. INTERACT, a transformer-based model, accurately predicts cell-type-specific DNA methylation profiles in the human brain with an average AUC of 0.99, identifying regulatory variants and enhancing fine mapping for disorders such as schizophrenia, depression, and Alzheimer's disease [45]. Similarly, scMeFormer excels in imputing DNA methylation states in single cells, achieving high-fidelity imputation even at 10% CpG coverage and uncovering thousands of schizophrenia-associated DMRs in prefrontal cortex datasets [46]. Despite these breakthroughs, single-cell DNA methylation profiling faces significant challenges, including high technical noise due to low input material, stochastic coverage, and complex protocols, which compromise data reproducibility and reliability [38, 47]. The high costs and technical complexity further limit scalability for large cohorts, particularly in blood-based applications. Consequently, these limitations have hindered the clinical translation of single-cell methylation data. To date, no successful machine learning models leveraging single-cell DNA methylation data have been implemented for disease diagnosis or classification in clinical settings, rendering this approach currently unsuitable as a standard diagnostic tool due to its uncertainty, high costs, and procedural complexity.

Research indicates that each DNA methylation detection technique is associated with specific technical biases that can impact the accuracy and reliability of methylation data. If not corrected, these biases can propagate to ML models, leading to inaccurate predictions or reduced model performance. For instance, in array-based methods, such as Illumina Infinium, technical biases include probe-specific effects (where hybridization efficiency varies across probes) and array-specific effects (which can introduce batch effects). Additionally, in dual-color arrays, dye bias can affect the quantification of methylation levels. To mitigate these biases, normalization methods such as loess normalization for dye bias and subset quantile normalization for array-specific effects are commonly employed [48]. In sequencing-based methods, such as WGBS, technical biases include PCR amplification bias, where methylated and unmethylated sequences may be differentially amplified due to differences in GC content post-bisulfite conversion, potentially leading to over- or underestimation of methylation levels. Furthermore, incomplete bisulfite conversion can result in false-positive methylation signals. To address PCR bias, approaches such as using alternative polymerases or optimizing PCR conditions have been explored, though their success varies [49]. A notable method to correct PCR bias involves cubic polynomial regression, where DNA samples with defined methylation levels are analyzed in parallel to generate a regression curve, which is then used to correct biased methylation data from samples of interest, applicable across various loci without requiring locus-specific optimization [50]. Additionally, the TET-assisted pyridine borane sequencing (TAPS) technique offers a bisulfite-free alternative, converting 5mC and 5hmC into dihydrouracil and subsequently to thymine during PCR, reducing base composition bias and DNA damage while preserving longer sequences up to 10 kb, thus enhancing library complexity and sequencing quality [51]. Third-generation sequencing platforms, such as single-molecule real-time (SMRT) and nanopore sequencing, further mitigate biases by enabling direct methylation detection without PCR amplification, achieving read lengths up to 1 Mb and improving accuracy in methylation analysis [51]. These biases significantly affect ML model reliability, particularly in disease prediction, as distorted data can lead to poor generalization to new samples or misclassification, underscoring the importance of robust preprocessing and bias correction strategies.

Techniques like the Infinium Methylation 450K BeadChip and Infinium MethylationEPIC arrays are highly favored due to their extensive coverage of CpG sites and high accuracy, making them invaluable in epigenetic research and biomarker discovery, especially in cancer studies. Bisulfite sequencing further enhances detailed methylation mapping at single-base resolution despite its high cost and data analysis demands. These techniques advance the diagnosis of various diseases. However, the choice between multiple biochemical methods is largely limited to the purpose of the study and diagnosis (Fig. 2). For example, if target regions are unknown, genome-wide methylation profiling or searching for DMRs are appropriate choices. Techniques such as the Infinium Methylation 450K BeadChip, Infinium MethylationEPIC arrays, and WGBS are more commonly utilized than other methods due to their well-established pipelines and greater popularity in clinical studies. Additionally, these techniques benefit from having more accessible datasets compared to others [52]. Consequently, most clinical studies reviewed

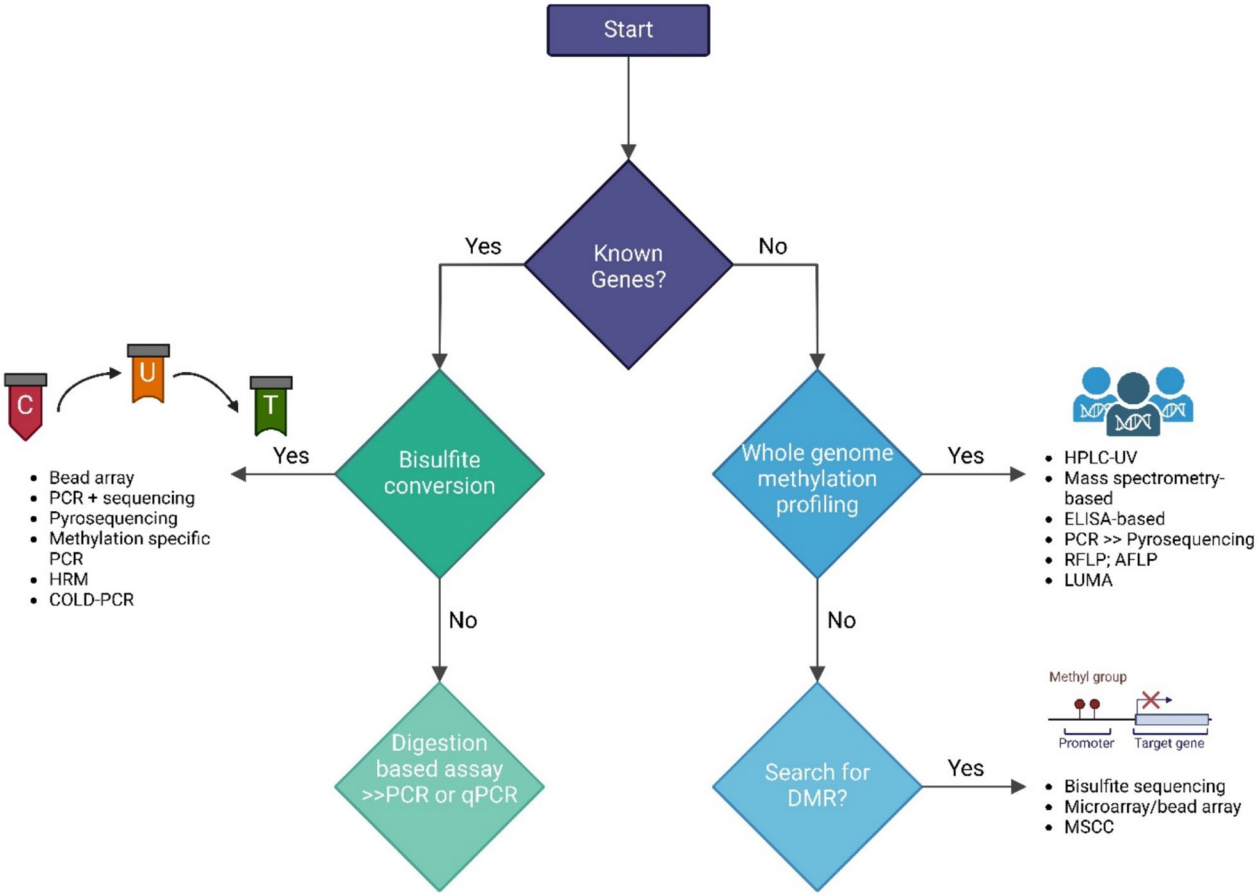


**Fig. 2** Decision flowchart for DNA methylation analysis method selection. Flowchart of a strategic guide for selecting appropriate DNA methylation analysis methods based on the specific requirements of the research. The process begins with the decision of whether the target genes are known. If yes, bisulfite conversion is recommended for detailed methylation analysis. If the genes are unknown, a choice is made between whole-genome methylation profiling or searching for differentially methylated regions (DMRs). Depending on the choice, various techniques such as HPLC–UV, mass spectrometry, ELISA-based assays, or more focused methods like bisulfite sequencing and bead arrays are employed. The chart also highlights the use of digestion-based assays for samples unsuitable for bisulfite conversion, providing a comprehensive pathway from sample preparation to specific analytical approaches tailored to different research goals

in this article have relied on the data outputs of these techniques for model training or as part of their clinical diagnostic approaches.

## ML approaches in epigenetics

Artificial intelligence (AI) is the intelligence in machines programmed to think and act like humans. According to Junaid et al., AI is designed systems with algorithms and computational processes that enable machines to gain and process information, make deductions, and perform other tasks [53]. ML is a branch of AI that creates systems that learn to identify patterns and predict based on the provided data [54]. ML includes various techniques such as supervised, unsupervised and reinforced learning. The supervised approach trains a model using a training dataset where labels have already been received in response to the input data. Unsupervised learning does not have labels and identifies relationships and patterns in the data. The reinforcement learning model trains an algorithm to face decisions like a game. In the end, the learning algorithm rewards positive decisions and punishes the negative ones; the algorithm does this alternately to achieve the final reward as much as possible. The models of ML achieve the goal of recognizing patterns and relationships that a human might not recognize when analyzing large data sets. ML has attracted increased attention over the past decade as a powerful tool for interpreting complex biological data in the field of epigenetics [55].

In the context of biomedical sciences, ML integrated into a workflow is a powerful tool for advancing diagnostics, personalized treatment strategies, and the discovery of hidden associative patterns in apparently complex biological processes (Fig. 3). The workflow starts with

defining the clinical or biological question and collecting relevant data, including patient records, DNA methylation data, and medical imaging. Preprocessing ensures data uniformity and suitability for analysis by normalizing, handling missing values, and encoding categorical variables. Feature selection and engineering refine the dataset to highlight impactful variables. The core involves selecting and training an ML model based on the data and problem specifics. Validation and testing assess the model's ability to generalize predictions using metrics like accuracy and precision. Results must align with clinical and biological knowledge to be meaningful and actionable. Continuous monitoring and updates are essential to maintain relevance and accuracy with new data and evolving scientific understanding.

#### Problem definition

A research problem describes a lack of knowledge or challenges current practices or ideas. It should be specific, clear, logical, relevant, and manageable. For example, a physician studying the epigenetic flows in the development of breast cancer may state the problem as finding DNA methylation changes that separate malignant from benign breast tissues [56]. Similarly, a biologist focusing on neurodevelopmental disorders may research specific methylation patterns on the AD genome: common in autism spectrum disorder [57]. This first step guides the subsequent stages; design experiments, sample selection, and methylation data analytics methods. Defining the research problem clearly is crucial as it guides the subsequent stages, including designing experiments, selecting samples, and choosing appropriate data analytics methods. The challenge lies in accurately defining the problem to ensure that the chosen samples and analysis methods are relevant and effective for the research goals [58].

#### Data collection

The data collection phase requires intricate planning and flawless execution, which commences with the careful selection of biological samples. Sample selection plays a critical role in capturing the methylation signature of the specific disease. Depending on the study, the choice of blood, tissue, or saliva samples determines the outcome, and it is essential to obtain samples that can reflect the methylation trend for that condition [59]. The second step in the process, size, is essential in developing prediction ML models on DNA methylation data because it directly affects the model's generalizability, statistical power, and risk of overfitting. Small data sizes can lead to a model that is fitted to the training data but captures the noise and not patterns that can be generalized, especially for complex models such as deep learning that require more data to prevent overfitting and perform stably [60–62]. Another issue is biological variability. Many case reports are biased and cannot represent the full picture, leading to the development of low-quality bias models. This affects the validation and testing by returning inflated results of the model's performance [60, 63]. To solve this problem in rare diseases, other methodologies can be applied, including data augmentation, transfer learning, cross-validation, regularization, and using simpler models. However, the solution is still obtaining a bigger dataset, which will better reflect reality because there will be more samples. The selection of assay type influences the quantity and the quality of the data which influences the outcome together with the adherence to ethical considerations and informed consent [64]. The key practice is to annotate data procured during the data collection phase with metadata that allows sophisticated correlation studies with clinical parameters [63].

#### Data preprocessing

Data preprocessing is a critical step in DNA methylation studies to ensure high-quality, reliable data for downstream analysis, with distinct approaches required for sequencing-based (e.g., WGBS, RRBS) and microarray-based (e.g., Illumina Infinium) methods due to their unique technical characteristics. For microarray-based methods, preprocessing begins with quality control to filter out low-quality probes or samples, addressing issues such as probe-specific effects (e.g., variable hybridization efficiency) and dye bias in dual-color arrays [65]. Normalization techniques, such as loess normalization for dye bias and subset quantile normalization for array-specific batch effects, is required to eliminate systematic variances caused, which ensures comparability of the sample metric, accurate methylation measurement utilizing background correction, and the calculation of beta values. Beta values are the quantitative measure of the CpG site's methylation state from fully unmethylated ($B = 0$) to fully methylated ($B = 1$) [66]. Additionally, every methylation matrix is plagued with missing data which needs to be addressed and should preserve the biological signal at the epigenetic level. Removing the data ultimately misplaces the information, and imputation or computation schema compensates for the lost data [67]. Missing data, often arising from failed probes or low signal intensity, are typically handled through imputation methods like k-nearest neighbors (kNN) to preserve biological signals without discarding valuable information. Batch effects in microarrays, stemming from variations in array lots, scanner settings, or chip-processing protocols, are identified using visualization tools like principal component analysis (PCA) and corrected with statistical methods such as the ComBat algorithm or linear mixed models treating batch as a random effect [68].

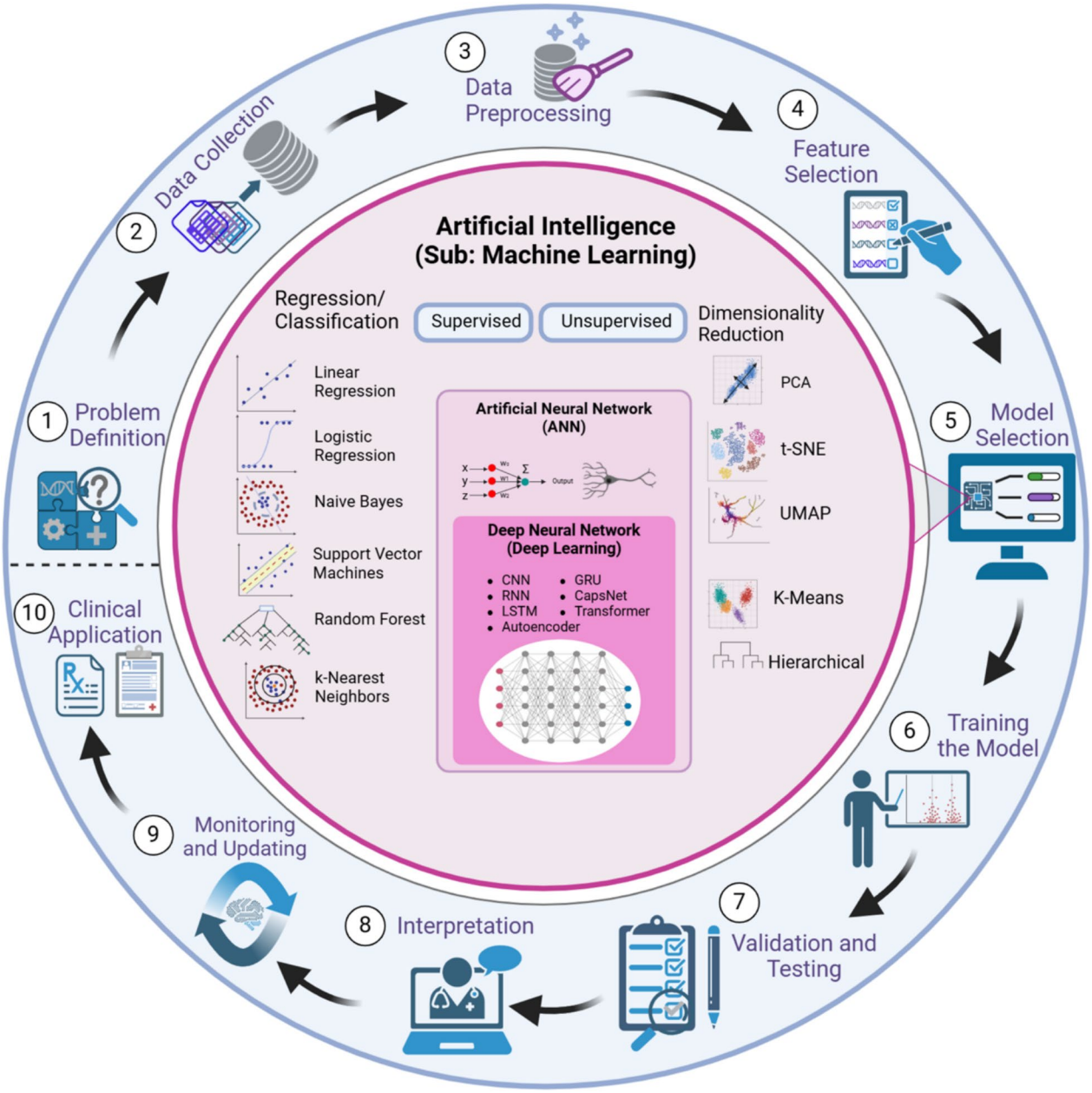


**Fig. 3** Comprehensive overview of the machine learning process, starting from the initial problem definition to the ongoing monitoring stages and final clinical application. It begins with the problem definition, where the specific challenge to be addressed using machine learning is identified. This is followed by data collection, where relevant data pertinent to the problem is gathered. Data preprocessing then takes place to clean and prepare the data, which may involve normalization, handling missing values, and transformation to make it suitable for analysis. The next step, feature selection, involves choosing the most relevant features that significantly contribute to the predictive accuracy of the model. Model selection follows, where an appropriate machine learning model is chosen based on the nature of the problem and the characteristics of the data. The model selection step is magnified in the green circle in three major methods in machine learning, including regression and classification (supervised), artificial neural network, and dimensionally reduction (unsupervised). Also, deep neural network methods are part of artificial neural networks. The selected model is then trained on the data to learn from it. Validation and testing are crucial stages where the model is evaluated on a separate dataset to assess its performance and generalizability. Once validated, the model undergoes interpretation where its results are analyzed to make data-driven decisions. Monitoring and updating ensure the model remain accurate and relevant by adjusting to new data or changes in the underlying data patterns. Finally, clinical application involves integrating the model into a clinical production environment to provide real-time predictions. This cycle reflects the iterative nature of machine learning, emphasizing continuous improvement and adaptation

In contrast, sequencing-based methods like WGBS and RRBS face distinct preprocessing challenges due to their reliance on bisulfite conversion and high-throughput sequencing. Quality control involves filtering low-quality reads and assessing bisulfite conversion efficiency, as incomplete conversion can lead to false-positive methylation signals. PCR amplification bias, where methylated and unmethylated sequences are differentially amplified due to GC content differences, is a significant issue that can distort methylation read ratios [49]. To address this, preprocessing pipelines may employ calibration with known methylation standards and correction methods like cubic polynomial regression to adjust for PCR bias, ensuring accurate quantification of methylation levels [50]. Normalization for sequencing data often involves adjusting for sequencing depth and coverage variability, with methods like quantile normalization or model-based approaches to account for technical variability. Batch effects in sequencing arise from differences in sample collection, DNA extraction, or sequencing runs, which can be mitigated by randomizing samples across sequencing lanes and applying statistical corrections like ComBat or surrogate variable analysis (SVA) [68]. The batch effect arises during multiple stages of the experiment and analysis. Variations in sample collection dates, handling, and DNA extraction process introduce batch effect. An example is when the samples are obtained on different dates, which might reflect the differences in environmental factors that influence sample quality. Storage conditions, including the preservative content, temperature, duration, and light exposure, affect DNA quality and theoretical quality, leading to discrepancies in the downstream methylation analysis [69]. Bisulfite sequencing can introduce errors during the bisulfite conversion, which can vary from batch to batch, affecting the percent conversion of C to U and affecting the critical metric. The laboratory environment is a source of variability due to different kind reagents. The platform on which the array is scanned could also have a prolific impact. Finally, the array or sequencing itself may be different in terms of library preparation processing protocol, scanner settings, array lots, and PCR amplification to induce measured batch effects. The data processor introduces variability from differences in software versions used in methylation data analysis [70].

#### Feature selection

Feature selection reduces the number of CpG sites from which informative markers can be selected, enhancing the precision and interpretability of predictive models [71]. DNA methylation data measures the methylation intensity of CpG sites. However, it is not mechanical that methylation states at CpG sites are independent by nature; they are spatially correlated, especially when they are in proximity [72]. Several factors could induce this phenomenon. One is that methylation and demethylation of adjacent sites are carried out by the same underlying biochemical processes and, hence, are likely to influence neighboring sites simultaneously [73]. Another explanation is that transcription factors that interact with DNA have the ability to alter methylation in entire regions [74]. These correlations in predictor space can lead to multicollinearity, interfering with the stability and reproducibility of ML algorithms. Multicollinearity refers to the inflation of the variance of estimated coefficients due to predictor variables that are highly correlated, making the model sensitive to small changes in the data and preventing robustness testing. Models with many correlated predictors tend to overfit the training data because they fail to adequately distinguish between signals and noise and consequently perform badly on unseen data [75]. Therefore, statistical corrections are required to achieve reliable findings. Feature selection identifies a small number of relevant CpG sites, which is necessary. Feature selection involves selecting a small subset of features by dropouts based on their importance to a defined task. For example, Doherty et al. mimic the process of measuring diversity downstream [76].

Biological context is important for feature selection in DNA methylation studies. For example, methylation patterns could differ between tissues or developmental stages. Feature selection eliminates irrelevant or redundant methylation settings, increasing the accuracy of predictive models and making the results more practically useful for clinical practice [71, 76]. Numerous statistical and ML methods are popularly applied, among which methylation levels should be the first consideration. Various statistical and ML techniques are employed for feature selection in DNA methylation studies. These include univariate methods such as t tests and ANOVA, which are used to identify differences in methylation levels between groups. However, it is crucial to consider the distribution of the data when applying these methods. DNA methylation data typically do not follow a normal distribution, often being binomially distributed, which necessitates specific transformations before applying standard statistical tests. Methylation levels at each probe, known as β-values, are calculated by comparing the signal from methylated probes to the total signal from both methylated and unmethylated probes. To address issues of normality and homoscedasticity, β-values are often transformed into M-values using a logarithmic transformation. This conversion is critical for applying statistical tests like t-tests and ANOVA, which assume a normal distribution. Moreover, to ensure the accuracy of the analysis, probes intersecting with single-nucleotide

polymorphisms and those located on the X or Y chromosomes are typically excluded. To prevent overfitting and enhance model simplicity, a crucial step in probe selection is performed on these M-values. Multivariable linear regression (MLR) modeling is commonly used in this process to select the most contributory probes for accurate classification while avoiding models that perfectly predict only the training set. Multivariate methods like LASSO (least absolute shrinkage and selection operator) and elastic net, which consider the joint effect of multiple CpG sites, as well as ML-based approaches such as random forests (RF) that can rank features based on their importance in classification or regression tasks, are also widely utilized [57, 77].

In addition to traditional statistical approaches for feature selection in DNA methylation studies, machine learning-based methods, such as mutual information (MI) and support vector regression (SVR)-based filter techniques, have gained prominence due to their ability to capture complex relationships in high-dimensional epigenomic datasets. MI quantifies the shared information between methylation features (e.g., CpG sites) and target variables, effectively identifying nonlinear dependencies, as demonstrated in a study on prostate cancer methylomes, where MI revealed distinct DNA methylation with RNA and copy number aberration associations between localized and metastatic stages [78]. Conversely, SVR-based methods leverage regression models to assign weights to features, prioritizing those with high predictive power. For instance, MethylCIBERSORT employs SVR for robust deconvolution, discarding markers with low reconstruction error, while ARIC enhances SVR-based feature selection by eliminating redundant markers using condition numbers to assess collinearity [79]. These approaches enable more precise identification of biologically relevant methylation markers, enhancing downstream analyses in epigenetic research.

#### Model selection

The type of model selected is contingent upon the nature of the research question, specifically whether the researcher is searching for associations between methylation patterns and disease status, predicting clinical outcomes based on methylation profiles, or understanding the cascade of events that regulate gene expression. For instance, if a researcher is looking to solve a binary classification task to determine healthy versus diseased, they may select logistic regression; should they seek to evaluate the relationship between methylation and quantitative traits, they might choose linear regression [80, 81] (Table 2). A recent study used MLR to predict local DNA methylation disorder in acute myeloid leukemia (AML) patients with DNMT3A-destabilizing variants, revealing significant associations with epigenetic heterogeneity after adjusting for mutation status, age, and gender [82]. In addition to the properties of the terminator dataset, such as dimensionality, distribution, and collinearity, model selection is often driven by whether the relationships being evaluated are linear. If the data are complex, high-dimensional, and involve interaction, more advanced models such as support vector machines (SVM), RFs, and neural networks may be better suited to identifying complex patterns and interactions in the methylation data. Model performance is commonly evaluated using various metrics. For classification tasks, metrics like accuracy, sensitivity, and specificity are used, while for regression tasks, mean squared error is often assessed. Cross-validation is a technique used to assess a model's performance by splitting the data into multiple subsets, training the model on some subsets while testing it on others. This approach helps in evaluating the model's ability to predict new, unseen data, detecting overfitting, and ensuring the model can generalize well to other datasets [83]. Furthermore, the choice of the best model is often guided by expert opinion, ensuring that the integration of biological information into the statistical analysis makes the findings more relevant.

It is evident that SVM especially works well in classification, but in the context of DNA methylation studies, it is best deployed in a dataset where a large margin exists between classes since this ML model is efficient in nonlinear relationships, thanks to its kernel functions. It can be used to learn the differences between methylation patterns that define different biological states or disease processes [84]. A recent study utilized a support vector machine (SVM) classifier within the EpiSign assay to analyze genome-wide DNA methylation profiles from the EPIC array, confirming a SETD2-related episignature in a patient with a novel syndromic multiple tumor phenotype, including sacral osteoblastoma, benign femoral bone tumor, peritoneal pseudomyxoma, and hypophyseal macroadenoma, achieving high concordance (MVP score > 0.5) with the Luscan-Lumish syndrome methylation profile for precise diagnosis [85]. SVMs have a unique advantage, which is that they are relatively stable in a high-dimensional state and avoid overfitting due to the regularization process. This means that the number of features can be much larger than the number of samples [57]. Although it can deliver high classification accuracy, it is less interpretable and requires much expert knowledge to discern the differences between the features that classify different classes, which makes it a challenge to extract biological insights directly from the model [86]. RF, on the other hand, is a family of ML models that are robust for high interactions and heterogeneity. It trains multiple decision trees during learning and combines the classification results by taking

**Table 2** Overview of conventional machine learning model methods in DNA methylation analysis

| Model | Application | Advantages | Limitations |
|---|---|---|---|
| Linear Regression | Predicts continuous methylation levels or disorder metrics, modeling linear relationships in epigenetic data | Simple, interpretable, and computationally efficient for continuous methylation outcomes with linear patterns | Assumes linear relationships, less effective for complex, nonlinear methylation patterns, sensitive to outliers |
| LASSO Regression | Performs feature selection and predicts methylation levels in high-dimensional datasets, reducing model complexity | Selects key CpG sites, reduces overfitting in high-dimensional data, provides interpretable coefficients | Assumes linear relationships, may discard correlated CpG sites, requires careful tuning of regularization parameter |
| Logistic Regression | Classifies methylation profiles for binary or multiclass disease diagnosis, modeling probability-based outcomes | Interpretable coefficients, effective for binary/multiclass classification, computationally efficient with regularization | Limited to linear decision boundaries, struggles with high-dimensional, nonlinear methylation data without preprocessing |
| Naïve Bayes | Classifies methylation patterns for tumor subtyping, leveraging probabilistic models for high-dimensional data | Fast, handles high-dimensional data, effective for small datasets with clear probabilistic patterns | Assumes feature independence, which may not hold for correlated CpG sites, reducing accuracy in complex datasets |
| Support Vector Machine (SVM) | Classifies high-dimensional methylation data for disease diagnosis, using linear or nonlinear kernels to capture complex patterns | Robust for high-dimensional data, effective with linear or RBF kernels, excels in multiclass tasks with calibration | Computationally intensive, sensitive to kernel choice and hyperparameter tuning, less interpretable without post-processing |
| Random Forest (RF) | Classifies high-dimensional methylation data and identifies feature importance for tumor subtyping and disease diagnosis | Handles high-dimensional data, robust to noise, provides feature importance for interpretability, computationally efficient | May overfit without tuning, less effective for small datasets, requires calibration for accurate probability estimates |
| k-Nearest Neighbors (kNN) | Clusters methylation patterns for disease subtyping, leveraging local similarities in high-dimensional epigenetic data | Simple, intuitive, no training phase, effective for clustering methylation patterns with clear local structures | Sensitive to noise, k-value selection, and high-dimensionality; poor scalability for large methylation datasets |
| Gradient Boosting Machines (e.g., XGBoost) | Classifies methylation profiles and identifies nonlinear interactions for disease prediction in high-dimensional datasets | Captures nonlinear patterns, robust to missing data, provides feature importance, high accuracy with tuning | Computationally intensive, sensitive to hyperparameter settings, prone to overfitting without proper tuning |
| PCA | Reduces dimensionality of methylation data for visualization and clustering, preserving variance in epigenetic profiles | Computationally efficient, preserves variance, simplifies visualization and clustering of methylation data | Assumes linear relationships, may lose nonlinear patterns, less effective for complex methylation datasets |
| t-SNE | Visualizes high-dimensional methylation data, capturing nonlinear structures for tumor subtype clustering | Effective for visualizing complex methylation patterns, captures nonlinear structures in low-dimensional space | Sensitive to parameter settings (e.g., perplexity), computationally intensive, non-deterministic, challenging for small datasets |
| UMAP | Reduces dimensionality and visualizes methylation data, preserving local and global structures for disease subtyping | Preserves local and global data structures, faster than t-SNE, effective for clustering and visualization | Sensitive to parameter tuning (e.g., neighbors), computationally demanding, less interpretable for complex datasets |
| K-Means | Clusters methylation profiles to identify disease or tumor subtypes based on epigenetic similarities | Simple, fast, effective for identifying methylation-based subgroups with clear cluster boundaries | Assumes spherical clusters, sensitive to initialization and noise, requires predefined cluster number |
| Hierarchical Clustering | Groups methylation profiles hierarchically to reveal disease or tumor subtype relationships in epigenetic data | Does not require predefined cluster numbers, captures hierarchical relationships in methylation patterns | Computationally intensive for large datasets, sensitive to noise, less robust for high-dimensional methylation data |
| Artificial Neural Network (ANN) | Captures nonlinear methylation patterns for disease classification and regression in high-dimensional datasets | Versatile for classification and regression, captures complex nonlinear patterns, scalable with sufficient data | Requires large datasets, computationally intensive, less interpretable without post hoc methods like SHAP |

**Table 2** (continued)

| Model | Application | Advantages | Limitations |
|---|---|---|---|
| Convolutional Neural Network (CNN) | Extracts local methylation features for tumor classification, leveraging structured epigenetic data patterns | Effective for structured data, extracts local features, achieves high accuracy in tumor classification | Requires large datasets, complex architecture design, limited interpretability without visualization tools |
| Recurrent Neural Network (RNN) | Models sequential methylation changes, capturing temporal dependencies in longitudinal epigenetic studies | Captures sequential dependencies, suitable for time-series methylation data analysis | Prone to vanishing gradients, computationally intensive, requires large sequential datasets for training |
| Long Short-Term Memory (LSTM) | Analyzes sequential methylation data for disease progression prediction, handling long-term dependencies | Robust for time-series analysis, handles long-term dependencies in sequential methylation data | High computational cost, complex architecture, requires careful tuning and large datasets |
| Capsule Neural Networks (CapsNets) | Classifies structured methylation data, preserving spatial relationships for improved tumor subtyping | Preserves spatial hierarchies, addresses CNN limitations, effective for structured methylation data | Computationally complex, limited adoption in methylation studies, requires large datasets for training |
| Gated Recurrent Unit (GRU) | Models sequential methylation patterns for disease classification, offering efficiency in time-series analysis | Efficient alternative to LSTM, captures sequential dependencies, lower memory usage | Less effective for very long sequences, requires large datasets, complex to interpret |
| Autoencoder (AE) | Reduces dimensionality and extracts features from high-dimensional methylation data for clustering or classification | Effective for feature extraction and noise reduction, supports unsupervised learning in methylation analysis | Requires large datasets, limited interpretability, reconstruction errors may affect downstream tasks |
| Transformer | Captures contextual relationships in methylation data for complex disease classification and tumor purity estimation | Excels in complex analyses via attention mechanisms, captures contextual epigenetic relationships | High computational cost, requires large datasets, interpretability challenges without post hoc methods |

a mode of the responses, making it possibly powerful for DNA methylation data with strong interactions [87, 88]. A recent study developed an RF-based classifier for tumors of unknown origin, accurately identifying the primary site in 97% of samples with 85% receiving high-confidence probability scores (≥0.9), demonstrating its reliability in guiding treatment decisions [89]. Similarly, another study employed the Heidelberg CNS Tumor Methylation Classifier (v12.8), an RF-based model trained on 7,495 methylation profiles, achieving 95% subclass-level accuracy across 184 CNS tumor subclasses, enhancing diagnostic precision in neuro-oncology [90]. While RF can easily handle co-linearity and high-dimensional data, the extent of feature importance allows biological interpretation as to which sites are critical in methylation. RF handles missing values and avoids overfitting [87, 91]. However, the high number of trees can be computationally intensive, and the features are challenging to interpret and how each feature contributes to predictions [92].

LASSO and ElasticNet are regularized regression methods widely used in clinical DNA methylation studies to address high-dimensional datasets, where the number of CpG sites exceeds the number of samples. These methods excel in feature selection and modeling associations between methylation patterns and clinical outcomes like disease status, prognosis, or treatment response. LASSO's L1 penalty shrinks less relevant coefficients to zero, identifying sparse, biologically relevant CpG sites, while ElasticNet's combination of L1 and L2 penalties handles correlated features, common in methylation data [52]. For example, a recent study used LASSO to develop a prognostic model for endometrial cancer recurrence, achieving AUC values of 0.671–0.725 [93], while another current research team applied ElasticNet to identify 27 methylation markers for stage II colon cancer recurrence risk stratification (AUC: 0.66–0.72) [94]. Both models reduce overfitting and enhance interpretability, making them ideal for sparse data and biomarker discovery [38]. However, LASSO's popularity is tempered by its limitations, as it may select only one feature from correlated groups, leading to unstable performance, and both methods assume linear relationships, missing complex patterns DNA methylation data [95].

Gradient boosting machines, particularly XGBoost, are increasingly popular in clinical DNA methylation studies for their ability to handle high-dimensional datasets with complex, nonlinear relationships between CpG sites and clinical outcomes like disease classification, prognosis, or treatment response [96]. For example, a recent study used XGBoost to validate 17 novel epigenetic biomarkers for anxiety disorders, achieving an AUC of 0.876, demonstrating its better diagnostic potential rather than RF in this study [97]. Additionally, XGBoost was integral to the NANOME pipeline, enhancing DNA methylation detection precision by 11% in nanopore long-read sequencing data for human B-lymphocyte cell lines [98]. XGBoost excels in managing missing data, offers interpretable feature importance scores, and provides high predictive power. However, its computational expense and risk of overfitting if not properly tuned are notable limitations, particularly in methylation analysis where noisy data or small sample sizes can exacerbate overfitting, and the complexity of tuning hyperparameters may hinder application in resource-limited settings [38].

Unsupervised learning methods, such as PCA, t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), K-means, and hierarchical clustering, are extensively employed in clinical DNA methylation studies to explore high-dimensional datasets and uncover hidden patterns without predefined labels. These methods are particularly valuable for dimensionality reduction and clustering to identify methylation-based subgroups or biomarkers associated with clinical outcomes like disease subtypes or progression. For example, a recent study used hierarchical clustering and K-Means on breast cancer methylation data from the Melbourne Collaborative Cohort Study and TCGA, revealing distinct tumor subtypes, though agreement between clustering methods varied (ARI < 0.7), highlighting the impact of algorithm and parameter choices [99]. Similarly, another current research team applied PCA to methylation array data from cutaneous melanoma, identifying biologically distinct subgroups characterized by intermediately methylated CpGs, effectively separating malignant from benign lesions [100]. Additionally, a recent study introduced a discrete wavelet transform (DWT)-based dimensionality reduction technique, outperforming PCA and UMAP in preserving spatial information for cancer classification, improving SVM accuracy while reducing computational demands [101]. These methods excel in discovering underlying structures, handling noisy data, and enabling visualization of complex methylation patterns, similar to the simplicity and intuitiveness of kNN. However, their reliance on unlabeled data complicates biological interpretation, necessitating follow-up analyses, and methods like t-SNE and UMAP are sensitive to noise and parameter settings, potentially yielding inconsistent results in small or noisy methylation datasets [52, 75, 95].

Neural networks, especially deep learning (DL) architectures, are suitable for modeling intricate nonlinear combinations, making them the best for capturing complex DNA methylation patterns. Deep and shallow learning models have the capability of learning from the classification and regression of data and transforming the raw data into discrete features, respectively. Although

sophisticated and require optimized power for learning, neural networks enjoy good featureability [92]. Artificial neural networks (ANNs), inspired by the human brain's interconnected neurons, capture nonlinear patterns through weighted layers; a recent study used ANNs to detect ovarian cancer via circulating cell-free DNA methylation, achieving an AUC of 0.99 [102]. Convolutional neural networks (CNNs), designed to extract local features through convolutional and pooling layers, are effective for structured methylation data; a current research team applied CNNs to classify papillary thyroid carcinoma subtypes, achieving an AUC of 0.91 [103]. Deep autoencoders (AEs), comprising encoder and decoder components, reduce high-dimensional methylation data for feature extraction; a recent study utilized a variational autoencoder (VAE) in RareNet to classify rare cancers, achieving a 96% F1-score [104]. Capsule neural networks (CapsNets), using capsules to preserve spatial relationships, address CNN limitations. Recurrent neural networks (RNNs), with their feedback loops, are suited for sequential methylation data. Long short-term memory (LSTM) units, an advanced RNN variant with gates to manage long-term dependencies, enhance sequential data analysis; a recent study employed an LSTM-based MTMBLAE model to predict Alzheimer's disease progression, achieving an AUC of 0.9797 [105]. Gated recurrent units (GRUs), with simpler reset and update gates, offer efficiency over LSTMs; a current study used the EpiBrCan-Lite GRU model to classify breast cancer subtypes with 95.85% accuracy [106]. Transformer-based models, leveraging multi-head self-attention for contextual understanding, excel in complex methylation analyses; MethylBERT recently identified tumor-derived methylation patterns in liquid biopsies, achieving a correlation of 0.987 for tumor purity estimation [107]. These models not only provide high accuracy, but also enable feature extraction and pattern discovery, though they often require large datasets and significant computational resources for training [108].

To effectively utilize deep learning models in DNA methylation research, it is crucial to align the model choice with the research goals and data characteristics. Start with autoencoders to reduce dimensionality and identify patterns in high-dimensional methylation data, which is ideal for initial exploration or noise reduction. Use CNNs or RNNs if the study focuses on classifying disease states or predicting phenotypic traits based on methylation patterns. CNNs are particularly useful for spatially structured data, while RNNs, including LSTMs and GRUs, excel in sequential or temporal data, such as time-series methylation changes. For combining DNA methylation with other omics data like transcriptomics, and genomics, transformers provide a robust framework for integrating diverse datasets by leveraging their attention mechanisms to identify relevant interactions. Techniques like SHapley Additive exPlanations or gradient-weighted class activation mapping (Grad-CAM) can help interpret DL outputs. However, DL models require large datasets to train effectively without overfitting and can be opaque, making it difficult to interpret how the model arrives at its predictions. Additionally, the training process can be resource-intensive, requiring significant computational power and time [92].

Interpretability in machine learning models is essential for clinical DNA methylation studies, as it fosters trust and supports biomarker discovery by clarifying how models derive predictions from complex epigenetic data. Linear machine learning models, such as logistic regression or random forest, often provide clearer interpretability than nonlinear deep learning models due to their straightforward decision boundaries, which allow for intuitive connections between feature weights and biological insights in methylation data, as seen in models capturing linear relationships between CpG probes and tumor classes. SHapley Additive exPlanations (SHAP) enhances model interpretability by quantifying the contribution of each CpG site to the model's output, assigning importance scores that reveal which methylation features drive clinical predictions [96]. Grad-CAM improves interpretability by generating visual heatmaps that highlight critical genomic regions influencing model decisions [109], particularly useful for spatially structured methylation data. For example, a recent study applied SHAP to the Heidelberg brain tumor classifier, a random forest-based model, to identify key methylation probes distinguishing 82 tumor and 9 control classes, creating a 3D array of probe usage across 428,799 genomic sites [21]. This approach enabled the development of an interactive web application (shiny app) that visualizes class-specific methylation patterns, linking probes to functional regions like enhancers and CpG islands, thus aiding biomarker discovery and therapeutic target identification. While, another recent study utilized a perceptron-based neural network in the crossNN framework to classify over 170 tumor types, leveraging feature importance weights to identify critical CpG sites linked to gene promoters and enhancers as an interpretability strategy, achieving 99.1% precision for brain tumors and 97.8% for pan-cancer classification [110]. However, this model emphasized simpler linear relationships for enhanced interpretability. These methods enhance transparency, connecting methylation patterns to clinical outcomes like disease diagnosis, though their effectiveness can be limited by model complexity and data noise.

Up to this point in this review, we have introduced the most recognized machine learning models for clinical

DNA methylation analysis, as summarized in Table 2, highlighting the need for benchmarking to navigate the complexity of selecting an optimal model for accurate and reliable diagnostic outcomes. A comprehensive evaluation of machine learning workflows on a brain tumor 450k DNA methylation cohort (2,801 samples, 91 classes) compared RF, Elastic Net, SVM, and XGBoost, revealing distinct performance profiles [75]. RF with default settings achieved a misclassification error (ME) of 4.8% and AUC of 99.9%, while tuned RF with multinomial ridge-penalized regression calibration reached a lower ME of 2.7% and Brier Score of 0.046, benefiting from its ability to handle high-dimensional data efficiently in under 40 min [75]. Elastic Net, particularly with a 10k probe feature space, outperformed uncalibrated classifiers with an ME of 2.7%, Brier score of 0.048, and AUC of 99.9%, showcasing its strength in managing sparse methylation data [75]. SVM with linear kernels, when calibrated with MR, achieved the lowest ME (2.1%) and Brier Score (0.039), excelling in multiclass classification tasks due to its robust margin optimization [75]. However, XGBoost with default settings underperformed (ME ~ 16%), though tuned versions improved to an ME of 5.1% and Brier score of 0.15, indicating sensitivity to hyperparameter optimization [75]. Another study on whole-blood DNA methylation for Parkinson's disease and schizophrenia classification compared logistic regression, SVM, XGBoost, CatBoost, LightGBM, attentive interpretable tabular learning (TabNet), and neural oblivious decision ensembles (NODE), with tree-based models like LightGBM achieving up to 97% accuracy on harmonized Parkinson's data, highlighting the impact of data harmonization on performance [96]. In contrast, a study on heart failure with preserved ejection fraction risk prediction using 25 CpG sites and five clinical features found the factorization-machine-based neural network model (AUC 0.90) outperforming benchmark models like RF (AUC 0.63–0.83) and Cox regression (C statistic 0.85), due to its ability to integrate multiomics data effectively [111]. A depression classification study across eight cohorts ($n = 1942$) showed RF achieving the highest AUC (0.76) on non-harmonized data with limma-selected features, while deep learning models like joint fully-connected autoencoder-classifier reached AUCs up to 0.91 with pre-selected features, though harmonized data reduced performance (AUC < 0.57) [112]. These findings underscore that no single model universally excels; RF and ELNET offer computational efficiency and robust performance for high-dimensional methylation data, while SVMs shine in calibrated multiclass settings, and deep learning models like factorization-machine-based neural network model excel in multiomics integration. However, model performance heavily depends on data characteristics (e.g., harmonization, feature selection) and diagnostic goals (e.g., multiclass vs. binary classification). Researchers should conduct dataset-specific benchmarking, testing multiple models with tailored preprocessing (e.g., harmonization to reduce batch effects, feature selection via limma or variance-based methods) and calibration strategies (e.g., multinomial ridge-penalized regression or Platt scaling) to optimize performance. Cross-validation schemes, such as 5 × fivefold nested CV, and evaluation metrics like AUC, BS, and log-loss should guide model selection, while interpretability tools like SHAP can enhance clinical trust. Ultimately, the choice of model should align with the specific methylation dataset and clinical objective, often requiring iterative benchmarking by researchers to ensure robust, reproducible results.

#### Training the model

To analyze the methylation dataset, the researchers split it into three sets: training, validation, and test. As mentioned above, the training set is utilized to train the model, the validation set is necessary to adjust and optimize the hyperparameters, thus preventing overfitting, and the test set would be used to measure the model's performance on new and unseen data [61]. During the hyperparameter tuning phase, strategies such as grid search, random search, and particularly nested cross-validation can be employed to optimize model performance while minimizing the risk of overfitting, especially when working with small sample sizes. Nested cross-validation incorporates an inner loop for hyperparameter tuning and an outer loop for model evaluation, which helps provide unbiased performance estimates and improves the generalizability of the final model [113].

When training the model, it is taught and learned by minimizing the error between the measured and predicted outcomes to the underlying relationships and patterns between the DNA methylation features and the outcome of interest [114]. When the model is evaluated and calibrated, the imperative hyperparameters are adjusted and fine-tuned to get the best validation set. Afterward, the hyperparameters of the model are adjusted and validated using the validation set. This step is premised on finding the best model.

The training process requires robust computational platforms and programming frameworks tailored to the unique characteristics of DNA methylation data. Python has emerged as a popular choice due to its versatile libraries. For deep learning, frameworks like TensorFlow and PyTorch are preferred for their scalability and performance in handling high-dimensional data [115, 116]. Meanwhile, scikit-learn offers accessible tools for traditional machine learning methods, enabling efficient feature selection and

classification [117]. For researchers favoring R, packages such as caret and mlr3 are widely utilized, providing extensive capabilities for model training, validation, and visualization [118, 119]. Recently, the integration of AutoML frameworks has enabled these approaches to be automated and optimized, thereby reducing the need for manual intervention. This streamlining not only enhances reproducibility and efficiency, but also lays the groundwork for the development of robust tools that can be translated into clinical applications.

Deep learning approaches in particular benefit from computational acceleration through GPUs (graphics processing units) or TPUs (tensor processing units). These resources are critical for efficiently processing large-scale datasets like whole-genome methylation profiles, significantly reducing training time and improving performance [120, 121]. Traditional machine learning methods are generally computationally less intensive and can be executed effectively on standard CPUs (central processing units) [122]. For larger datasets or more complex tasks, multicore CPUs or small-scale parallel processing can further enhance performance. Cloud-based platforms, such as Google Colab and AWS EC2, provide researchers with scalable access to these resources, democratizing the use of model training for methylation data analysis.

Moreover, the selection of training frameworks and computational resources must align with the goals of the study. Curated datasets and pretrained models available from platforms like TensorFlow Hub and Bioconductor facilitate a smoother training pipeline, allowing researchers to focus on domain-specific challenges [123]. The iterative process of model training and hyperparameter tuning, combined with these tools, enables researchers to derive meaningful insights from complex DNA methylation datasets.

#### Model validation and testing

After training is complete, the model's performance is evaluated on an independent test set to obtain an unbiased estimate of the predictive accuracy and generalization ability of the model. The performance of the model is commonly evaluated using metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC), among others [114]. However, the unique characteristics of DNA methylation data introduce specific confounding factors such as cell-type composition, spatial correlation of CpG sites, and methylation drift that must be addressed during validation to achieve robust and clinically relevant predictions. Additionally, detecting whether the model has learned spurious patterns driven by these confounders rather than biologically meaningful signals is critical for model reliability.

Cell-type composition is a major confounder, as tissues often comprise heterogeneous cell populations with distinct methylation profiles. Failure to account for cellular heterogeneity during validation can lead to models that capture cell-type-driven signals rather than disease-specific patterns, resulting in inflated performance metrics. To address this, preprocessing steps like cell-type deconvolution (e.g., using EpiDISH or RefFreeCellMix) should be validated in the test set to confirm consistency [124, 125]. To detect spurious patterns, sensitivity analysis can be performed by comparing model performance with and without cell-type correction. A significant drop in performance without correction suggests reliance on cell-type-driven patterns.

Spatial correlation of CpG sites can lead to overfitting if feature selection does not account for correlated methylation levels among neighboring CpG sites [126]. During validation, feature importance analysis using methods like SHAP values or permutation importance can reveal whether the model relies on biologically irrelevant CpG sites driven by spatial correlations [96]. For example, if highly correlated CpG sites dominate feature importance but lack known biological relevance, the model may have learned spurious patterns. Techniques like CpG clustering in preprocessing can mitigate this, but their effectiveness must be confirmed in the test phase [127].

Similarly, methylation drift, driven by age or environmental exposures, can lead to models learning patterns unrelated to the disease. A model with high AUC on both training and test sets but strong correlation with age (e.g., via Pearson correlation of predictions with age metadata) likely relies on age-related methylation drift. To diagnose this, residual confounder analysis can assess whether model predictions correlate with age or other confounders. For example, a high correlation (e.g., $|r| > 0.5$) suggests that the model has learned spurious patterns, necessitating preprocessing adjustments like age-normalized methylation profiles or regression adjustment for age [128]. Additionally, external validation with cohorts of diverse age distributions can confirm whether the model generalizes or fails due to age-driven patterns.

Other confounders, such as genetic variation (e.g., SNPs or mQTLs), demographic factors (e.g., age, sex, ethnicity), lifestyle factors, disease status, batch effects, and technology/platform variability, can also lead to spurious patterns [129]. For instance, batch effects, even if addressed in preprocessing using methods like ComBat [43], may persist as residual noise in external datasets. To detect this, visualization techniques like PCA or UMAP can be applied to test set data to check for batch-driven clustering. Similarly, domain knowledge integration, comparing selected CpG sites to known disease-related

epigenetic markers, may reveal whether the model has learned biologically implausible patterns.

The model can subsequently be refined iteratively using various algorithms, feature sets, or preprocessing methodologies to achieve further performance optimization. This may involve modifying the model architecture for neural networks, the feature selection methods, or the preprocessing methodologies employed based on the experiment insights obtained after model evaluation. Researchers may also opt to externally validate the model using data from a separate cohort or experiment. External validation is critical, as it helps improve the validity and generalizability of the model through corroboration by different data across populations, between experiments, or across data sources [130]. It is important to note that an ML model can only perform as well as there are actual underlying patterns in the data. For example, increasing the number of hyperparameter configurations, changing feature selection methods, or altering the ML model cannot improve the performance of a classifier if the cases and controls have no clear underlying differences. This is because ML algorithms cannot create meaningful predictions from data with insignificant underlying patterns and may only end up mapping random patterns in the data. This underscores the importance of ensuring that the data used in the training and validation of models actually contain biologically relevant and statistically significant differences to improve the ability to generate accurate and robust prediction models [60–63].

#### Interpretation

The interpretation phase involves unravelling the complex relationships between DNA methylation patterns and various biological or clinical outcomes and translating these findings into actionable knowledge that can inform clinical practice and advance scientific understanding. The most contributing methylation sites or genomic spots can be identified by evaluating the coefficients or feature importance scores of supervised models. These biomarkers may be diagnostic, prognostic, or anticipatory markers for disease advancement, development, or treatment response [114]. The interpretation also entails aligning biomarkers with the broader biological frameworks of pathways, regulatory brokers, and cellular processes underpinning the results. That is, identifying how methylation correlates with gene expression changes, chromatin configurations, and cellular phenotypes, which might provide a hint on the main processes behind disease pathophysiology or treatment response [131]. Ultimately, the goal of the interpretation is to translate epigenetic assessments into clinical plans.

#### Monitoring and updating

Continuous monitoring and updating of DNA methylation-based models are crucial for maintaining accuracy and relevance in disease diagnosis. This process ensures that models adapt to new findings and remain effective in clinical settings. Recent studies emphasize the importance of incorporating interaction and nonlinear effects between CpGs to capture complex relationships. For example, DNA methylation has shown additional predictive benefits for 10-year Type 2 diabetes risk beyond standard risk factors like age, sex, and BMI. This was demonstrated using large datasets with genome-wide DNA methylation and electronic health records, with results validated in another cohort [132]. The best-performing updated model with the Generation Scotland cohort DNA methylation data, achieved an AUC of 0.872, compared to 0.839 for models using only standard risk factors, and the area under the precision-recall curve improved from 0.227 to 0.302. These improvements demonstrate the substantial benefit of integrating epigenetic data and retraining models with updated inputs. Similarly, the EAGLING model [133] improves the prediction of methylation levels at unmeasured CpG sites by using information from nearby CpGs. Three main changes led to better performance. First, the model no longer uses DNA sequence features, which makes it significantly faster without reducing accuracy. Second, it uses the methylation value of the closest CpG site instead of combining values from neighboring sites, which improves precision. Third, it was trained on a larger dataset, increasing the number of tissues and cell lines from 14 to 33. These changes led to measurable improvements. The model's accuracy increased from 87.11 to 89.65%. The AUC improved from 0.8595 to 0.8645. The correlation coefficient and concordance also increased, from 0.8321 to 0.8441 and from 0.8375 to 0.8532, respectively. These results show that simplifying the model and expanding the training data can improve both speed and prediction quality. These studies highlight the dynamic nature of methylation-based predictive models and their iterative refinement to enhance predictive capabilities. As new data is acquired, it is essential to re-run feature selection and retrain models to stay current with the most relevant features, thereby improving accuracy and predictive power.

## Clinical applications of ML in epigenetics

The clinical application of DNA methylation biomarkers holds significant promise across various medical fields. In cancer, these biomarkers aid in early detection, prognosis, and personalized treatment strategies. For neurodevelopmental disorders, methylation profiles enhance our

understanding of epigenetic mechanisms and improve diagnostic accuracy. In multifactorial diseases, such as cardiovascular conditions and dementia, they provide insights into the interplay of genetic and environmental factors, aiding in risk prediction and personalized management strategies. As research continues to evolve, it is expected that more methylation-based diagnostic tools and therapies will be used widely to improve patient care and outcomes.

### Prediction

DNA methylation, a key epigenetic modification, plays a pivotal role in disease development by influencing gene expression and genomic stability. Its patterns, characterized by global hypomethylation and gene-specific hypermethylation, are extensively studied in various diseases, particularly cancer, neurological disorders, cardiovascular diseases, and rare genetic syndromes. This section explores the integration of ML with DNA methylation profiling to enhance diagnostic and prognostic capabilities, organized by disease category for clarity and coherence (Table 3). This section synthesizes key advancements in predictive models, organized by disease category, with critical comparisons of methods, discussion of unresolved challenges, and actionable guidelines for clinical translation. These predictive insights lay the foundation for diagnostic, therapeutic, and monitoring applications detailed in Sects. "Diagnosis"–"Follow-up clinical parameters".

#### *Cancer*

DNA methylation changes are closely associated with the development of cancer and represent one of the most extensively examined epigenetic changes in oncology. In neoplasms, the DNA methylation pattern undergoes significant changes, characterized by global hypomethylation and gene-specific hypermethylation [91]. Hypomethylation in the neoplastic cells is responsible for genomic instability and activation of proto-oncogenes. The hypermethylation of CpG islands in promoter regions is a cause of the inactivation of tumor suppressor genes, promoting the development of tumors and the failure of apoptosis. These epigenetic alterations are crucial for the initiation and progression of cancer and affect cellular processes such as DNA repair, cell cycle regulation, and apoptosis. Esteller's review has also showcased that DNA methylation plays a double stand in oncogenesis and is a potential marker for early detection, prognosis, and response prediction [134].

Recent studies involving the complexity of cancer heterogeneity and metastasis applied advanced ML with DNA methylation profiling to help increase diagnostic accuracy and trace tumor etiology. One study analyzed methylation patterns across 24 cancer types using various algorithms that showed a 99% accuracy in classifying cancer types according to the site of origin and whether it is primary or metastatic [135]. Another study proposed a method for the use of cfDNA methylation profiles to detect early cancers based on a semi-reference-free deconvolution algorithm that showed a sensitivity of 86.1% for cancer early detection and an average accuracy of 76.9% for tumor localization at a specificity of 94.7% [136]. A novel multilayer perceptron tumor-type-specific hierarchical model using DNA methylation arrays data, indicated 99% in the test set and 93% in the external validation set in identifying the site and type of tissue within the 27 cancer types, which means it can be used for primary and secondary cancer identification [137]. Essentially, the results of the Genome-Derived-Diagnosis Ensemble, GDD-ENS, which uses deep neural networks on genomic data derived from a targeted cancer gene panel, may be useful in sorting the 38 different cancer types and offers an alternative to genome sequencing. In this research, the model accurately classified 94.46% of 1660 cases in the cancer genome atlas (TCGA) data set, and 91.4% of 677 cases in the gene expression omnibus (GEO) [138]. Finally, the study that looked into DNA methylome profiling versus the most common brain metastases types elucidated the BrainMETH classifier, which categorizes brain metastases into tissue of origin and subtype, showing the epigenetic-field-relevant understanding in the development of future treatments for metastatic cancers. This model consist of three classifiers with AUCs more than 90% to classify between primary and metastatic brain tumors, among brain metastases from different tumor of origin, and among breast cancer with brain metastases from different molecular subtypes [139].

The burgeoning integration of ML and DNA methylation markers into commercial platforms is revolutionizing cancer diagnostics, enabling earlier detection with greater precision. Among these innovations, Galleri®, developed by GRAIL, stands out as a multi-cancer early detection test. Employing WGBS, this test assesses methylation patterns in circulating cell-free DNA (cfDNA) from both cancerous and normal cells. Remarkably, its specificity for over 50 cancer types reached a signal detection rate of 99.5% [13]. Galleri® employs custom software that classifies samples based on methylation patterns specific to regions indicative of particular cancer types. This process involves a dual ensemble logistic regression approach: one regression determines the cancer/non-cancer status, while the other ascertains the tissue of origin among predefined sites [140]. However, its sensitivity as a performance metric has limitations; while the absolute number of detected cancers increases with the addition of cancer classes, the overall sensitivity might decrease. For instance, an overall sensitivity of

**Table 3** Summary of studies predict or classify human various diseases with machine learning trained models based on DNA methylation data

| Disease/Condition* | Model(s) | Data type | Sample size | Predicted performance | Ref |
|---|---|---|---|---|---|
| 24 Cancer Types | SVM, RF, XGBoost, Naïve Bayes | 27 K & 450 K array of tumor | 9303 (TCGA = 8840 & GEO = 463) | Average accuracy: 0.981<br>F1-scores: 0.931- 1.0 | [135] |
| Early 5 Cancer types Detection | Semi-reference-free Deconvolution Algorithm | 450 K array of blood cfDNA | 207 normal<br>113 late-stage<br>79 early-stage | Sensitivity: 86.1%, Accuracy: 76.9%),<br>Specificity: 94.7% | [136] |
| 27 Cancer Types | Multilayer Perceptron (Tumor-type-specific Hierarchical Model) | 450 K array of tumor | 7735 from TCGA & GEO | Accuracy: 99% (test set), 93% (external validation set) | [137] |
| 38 Cancer Types | GDD-ENS (Deep Neural Networks) | 450 K array of tumor | 1660 (TCGA), 677 (GEO) | Accuracy: 94.46% (TCGA), 91.4% (GEO) | [138] |
| Brain Metastases | Three RF Classifiers | 450 K array of tumor | 96(brain metastases)<br>1860(GEO)<br>165 (validation set) | AUC > 90% for classifying primary vs. metastatic, tumor of origin, and breast cancer subtypes | [139] |
| Over 50 Cancer Types | Galleri® (Dual Ensemble Logistic Regression) | WGBS of blood cfDNA | 15,254 (CCGA study) | Specificity: 99.5%, Sensitivity: 51.5% (overall), 76.3% (12 cancer classes) | [13, 140] |
| CNS Tumors | Heidelberg CNS Tumor Methylation Classifier V11 (RF) | 450 K, EPIC array, & WGBS of tumor | 2801 (91 classes) | AUC: 0.99<br>Sensitivity: 0.989<br>Specificity: 0.999 | [10] |
| CNS Tumors | Heidelberg CNS Tumor Methylation Classifier V12.8 (RF) | 450 K, EPIC, & EPICv2 array of tumor | 7495 (184 subclasses) | AUC of 184 > 75%<br>AUC of 175 > 90%<br>95% subclass-level accuracy, Brier score: 0.028 | [90] |
| Lung vs. Head/Neck Cancer | Artificial Neural Network | EPIC array of tumor | 279 patients | Accuracy: 96.4%<br>Prediction accuracy 99% < : 92.1% | [142] |
| Sarcoma | RF | 450 K & EPIC array of tumor | 1,077 (62 subtypes)<br>428 (validation set) | classifier prediction score ≥ 0.9: 75%<br>class matched with the institutional diagnosis: 61% | [143] |
| Sinonasal Tumors | SVM | 450 K & EPIC array of tumor | 429 sinonasal tumors<br>8104 other tumors & normal<br>52 (test set) | Accuracy in sinonasal validation cohort: 1.0,<br>Specificity: of 0.982 | [144] |
| Brain Tumors | RF | Nanopore WGS of tumor | 382 (46 subtypes) | Sensitivity: 80.4%, Specificity: 100%,<br>concordance Cross-lab validation: 90.9% | [145] |
| CNS Tumors (Intraoperative) | Sturgeon transfer-learned neural network | Nanopore WGS of tumor | 50 (Retrospective samples)<br>25 (during serguries samples) | Retrospective correct diagnosis: 90%<br>Correct diagnosis within 90 min during serguries: 72% | [146] |
| CNS Tumors | MethyLYZR (Naive Bayesian) | 450 K array & Nanopore WGS of tumor | 75 (Nanopore runs), 2801 (91 classes of previous study 450 K) | Accuracy: 94.52%,<br>Predicts 91.94% out of 82.67% of samples correctly | [147] |
| CNS Tumors | MNP-Flex (XGBoost) | 450 K, EPIC, EPICv2 array, WGBS, Nanopore WGS, & Rapid-CNS$^2$ of tumor | 58,625 (184 class from whole cohort classifiable), 301(Rapid-CNS$^2$) | Family_level accuracy: 99.6%,<br>Subclasses_level accuracy: 99.2% | [148] |

**Table 3** (continued)

| Disease/Condition* | Model(s) | Data type | Sample size | Predicted performance | Ref |
|---|---|---|---|---|---|
| AD, ALS, PD | Mixed-Linear Model | 450 K & EPIC array of blood | 5551(cases),<br>4343(controls) | Mixed linear model-based omics association AUC: 0.69<br>multi-component mixed linear model-based omics association excluding the target AUC: 0.68 | [149] |
| AD | Deep Neural Network | 450 K array + RNASeq of prefrontal cortex of brain | 74 (cases),<br>68 (controls) | Training accuracy: 0.832 Training accuracy: 0.823 | [150] |
| AD | Multi-task Deep Autoencoders (Convolutional + LSTM) | 450 k Array of blood | 649 from ADNI cohort | AUC: 0.996,<br>20% performance increase with unscaled beta values | [151] |
| SZ | AutoML with SVM | qMSP Assays of selected genes of blood | 30 (cases),<br>30 (controls) | AUC: 0.755 (95% CI: 0.636–0.862),<br>Average precision: 0.758 | [152] |
| Depression | RF, Joint Autoencoder-Classifier | 27 K, 450 K, & EPIC array of blood | 1942 (8 cohorts from GEO & PSY) | AUC: 0.73–0.76 | [112] |
| PD, SZ | LightGBM, CatBoost | 450 K array of blood | 1582 (PD),<br>1108 (SZ),<br>2398(controls) from GEO | PD accuracy: > 97% (LightGBM),<br>SZ accuracy: 72% (CatBoost) | [96] |
| Very Late-onset Schizophrenia-like Psychosis | XGBoost | 450 K array & WGBS of blood & brain | 1218 (SZ & AD patients from GEO) | SZ vs. VLOSLPAUC: 1.0,<br>SZ vs. AD AUC: 0.95 | [153] |
| degenerative disease | RF + Multi-1D CNN + ResNet | 27 K, 450 K, & EPIC array | 6667 (training set)<br>1668 (testing set) from GEO | Average AUC: 97.4%<br>Average precision: 0.974 | [154] |
| SZ | SPLS-DA (LASSO Regularization) | 450 k array of brain | 353 (cases),<br>322 (controls) | Positive predictive value: 80% | [155] |
| Major Depression | LASSO Penalized Regression | EPIC array of blood | 9873 from GS:SFHS | AUC: 0.58 | [156] |
| AD | Robust Linear Model | EPIC array of blood | 88 | CN-to-MCI conversion (estimate = −0.052, *P*-value = 0.0243), MCI-to-AD conversion (estimate = −0.031, *P*-value = 0.0283) | [157] |
| AD | ANN | EPIC array of cfDNA of blood | 26 (cases),<br>26 (controls) | AUC (95% CI): 0.99 (0.95–1.0),<br>Sensitivity: 94.5%,<br>Specificity: 94.5% | [158] |
| AD | Logistic Regression (MetaboAnalyst) | EPIC array of cfDNA of blood | 25 (cases),<br>23 (controls) | AUC training: 0.928, AUC validation: 0.942, Sensitivity: 100%, Specificity: 90% | [159] |
| Multiple Sclerosis | GradientBoostingClassifier, LightGBM | Low-coverage WGBS of cfDNA of blood | 75 patients | AUC: 0.70–0.82 (MS vs. controls),<br>AUC: 0.74 (disability progression) | [160] |
| Gestational Diabetes Mellitus | RF | 450 K & EPIC array of cord blood & placenta | 98 (cases),<br>98 (controls) from GEO &<br>5 (cases),<br>5 (controls) from clinical cohort) | AUC: 0.89 (public data), AUC: 0.82 (clinical cohort) | [161] |

**Table 3** (continued)

| Disease/Condition* | Model(s) | Data type | Sample size | Predicted performance | Ref |
|---|---|---|---|---|---|
| Diabetes | XGBoost | EPIC array of blood | 454 | AUC:0.816,<br>Accuracy: 0.745,<br>Precision: 0.764,<br>Sensitivity: 0.89<br>Hazard ratios: 2.86–4.08 (P < 0.001) | [162] |
| Coronary Heart Disease | LightGBM | 450 K array + RNASeq of blood | 2117 from Framingham Heart Study | AUC: 0.834,<br>Sensitivity: 0.672, Specificity: 0.864 | [163] |
| Coronary Heart Disease | RF | 450 K array + SNP array of blood | 1545 (training set)<br>from Framingham Heart Study,<br>142 (test set) | Accuracy: 78%, Sensitivity: 0.75,<br>Specificity: 0.80 | [164] |
| Heart Failure | LASSO + XGBoost + DeepFM (factorization-machine based neural network) | 450 K array of blood + electronic health record | 984 from Framingham Heart Study | AUC: 0.90 (95% CI: 0.88–0.92) | [111] |
| Osteoarthritis | Ridge, LASSO Regression | EPIC array of blood | 554 (training set from OABC),<br>183 (test set from OA & OAI) | AUC: 0.94 (radiographic),<br>AUC: 0.97 (pain),<br>AUC: 0.79 (combined),<br>AUC: 0.86 (any progression) | [165] |
| Multiple Phenotypes | LASSO, Elastic Net, Ridge Regression | EPIC array of blood | 831 from UCLA Health Biobank | AUC: 0.840 (95% CI: 0.807,0.871) | [166] |
| RCEM | Binary SVM Classifier | EPIC array of blood | 53 | Specificity: 98.32% | [168] |
| Neurodevelopmental Disorders | SVM | EPIC array of blood | 27(cases),<br>40(controls),<br>440 (validation controls) | Accurate classification of variants | [170] |
| Neurodevelopmental Disorders | EpiSign (SVM-based algorithms) | 450 K & EPIC array of blood | 258 (cases from 14 syndromes),<br>650 (controls) | Accuracy: 99.6%,<br>Sensitivity: 100%,<br>Specificity: 100% | [174] |
| BAFopathies | EpiSign (SVM-based algorithms) | 450 K & EPIC array of blood | 29 (cases),<br>156 (controls) | Accuracy: 98.8% | [175] |
| Neurodevelopmental Disorders | k-Nearest-Neighbor Classifier | EPIC array of blood | 57 NDD, 25 controls | Specificity: 100%, Sensitivity: 100% (*TRX*, *DNMT3A*, *KMT2D*, *NSD1*),<br>Sensitivity: < 40% (*CREBBP-RSTS*, *CHD8*),<br>Sensitivity: 70–100% (*KMT2A*, *KDM5C*, *CHD7*) | [178] |
| Developmental Disorders | SVM | Nanopore WGS | 20 patients | Accurate identification in 17/19 cases 89% | [181] |

**Table 3** (continued)

| Disease/Condition* | Model(s) | Data type | Sample size | Predicted performance | Ref |
|---|---|---|---|---|---|
| Developmental Disorders | SVM | Synthetic data based on EPIC & EPIC v2 array of blood | 169 synthetic (89 disorders) based on 4967 (53 studies from GEO), 128 cases & 461 controls (EpigenCentral validation), 184 cases & 461 controls (NSBEpi validation) | Accuracy EC: 0.997, Precision EC: 0.954, Sensitivity EC: 0.974, Specificity EC: 0.998, Accuracy NSBEpi: 0.972, Precision NSBEpi: 0.972, Sensitivity NSBEpi: 0.929, Specificity NSBEpi: 0.989, Average F1-score: 0.95 | [182] |

* This table summarizes machine learning models and their performance results for disease detection based on DNA methylation profiling, as described in Sect. "Prediction" of the article. It does not encompass all diseases detectable using this approach but highlights key and representative examples reported in recent studies. For a comprehensive overview, please refer to the main text and cited references

51.5% might detect more cancer cases in absolute terms compared to 76.3% sensitivity in a pre-specified set of 12 cancer classes. This nuance underscores the complexities of evaluating the clinical utility of multi-cancer detection tests [13].

DNA methylation-based classification has revolutionized molecular diagnostics in oncology, particularly in the precise categorization of central nervous system tumors, where it has been a cornerstone of the World Health Organization classification since 2021 [141]. The Heidelberg CNS Tumor Methylation Classifier (version 12.8), trained on an extensive dataset of 7495 methylation profiles covering 184 tumor subclasses, achieves a remarkable 95% subclass-level accuracy with a Brier score of 0.028, significantly outperforming traditional histopathological methods by reclassifying up to 12% of prospective cases and reducing inter-observer variability [10, 90]. This classifier, leveraging random forest algorithms and accessible via a free online platform (molecularneuropathology.org), enhances clinical decision-making by providing rapid, reproducible results without extensive onsite computational requirements. Beyond CNS tumors, methylation profiling has proven critical in resolving diagnostic challenges across other cancer types. For instance, an artificial neural network distinguishes primary lung squamous cell carcinomas from head and neck metastases with 96.4% accuracy in a 279-patient cohort, leveraging differential methylation patterns to inform treatment strategies [142]. Similarly, a sarcoma classifier trained on 1,077 methylation profiles accurately categorizes 62 subtypes, addressing morphological heterogeneity [143], while sinonasal tumor classification redefines poorly characterized entities like sinonasal undifferentiated carcinomas into four molecular classes, enabling targeted therapies based on epigenetic and mutational profiles [144]. Rapid advancements in nanopore sequencing have further accelerated diagnostics, achieving 80.4% sensitivity and 100% specificity for 46 brain tumor subtypes in just 21.1 h, with cross-lab validation confirming 90.9% concordance [145]. Intraoperative applications, such as the Sturgeon neural network and rapid nanopore sequencing, deliver real-time CNS tumor subclassification within 90 min, correctly diagnosing 72% of cases and minimizing neurological risks during surgery [146]. Nanopore sequencing has further revolutionized diagnostics, with the MethyLYZR naive Bayesian framework achieving 94.52% accuracy across 91 CNS tumor classes using just 7500 CpGs, delivering results within 15 min in a clinical setting [147]. The Rapid-CNS2 platform, validated on 301 samples, provides intraoperative methylation classification and copy number profiling within 30 min, achieving 99.6% accuracy for methylation families and 99.2% for classes across 78,000 samples [148]. These developments highlight the pivotal role of DNA methylation-based diagnostics in enabling precision oncology, with further applications in early detection, treatment planning.

#### *Neurological disorders*

Researchers are unlocking the molecular mysteries of neurological disorders by conducting advanced genome-wide DNA methylation studies and using ML techniques to identify common epigenetic signatures. In a comprehensive meta-analysis, researchers conducted a genome-wide DNA methylation study to identify shared methylation differences in the blood DNA of individuals with Alzheimer's disease (AD), amyotrophic lateral sclerosis (ALS), and Parkinson's disease (PD). Using a mixed-linear model that included both known and unknown confounding factors, they identified 12 genome-wide significant differentially methylated positions common across these disorders [149]. Another study built on this by integrating DNA methylation with gene expression data, selecting overlapping features at both the CpG sites and gene levels for AD model prediction training. Using a Bayesian method for optimal parameter selection, the study demonstrated that a deep neural network model with eight hidden layers containing 306 nodes, a learning rate of 0.02, and a dropout rate of 0.85 achieved superior performance. This integrated approach resulted in enhanced model accuracy compared to models using single-modal data like RF, SVM, and naïve Bayes [150]. Furthermore, another study developed two multi-tasks deep autoencoders: one based on a convolutional autoencoder and the other on a long short-term memory autoencoder. These models were designed to learn compressed feature representations by simultaneously minimizing reconstruction errors and maximizing the prediction accuracy of AD progression. When assessed on the Alzheimer's disease neuroimaging initiative cohort using the 450k array data, these multi-task deep autoencoders outperformed their single-task counterparts, achieving an AUC of 0.996. The study also noted that using unscaled beta values enhanced prediction performance by 20% in autoencoder methods, though this improvement was not seen in RF models. Additionally, increasing feature size led to longer training times and greater memory consumption [151]. In schizophrenia (SZ), an automated ML pipeline (AutoML with SVM) with targeted qMSP assays identified a five-feature biosignature, incorporating IGF2BP1, CENPI, PSME4, age, and sex, achieving an AUC of 0.755 (95% CI: 0.636–0.862) in 30 first-episode drug-naïve patients [152]. Additionally, depression studies across eight cohorts ($n = 1942$) using 12 ML strategies, including RF and joint autoencoder-classifier models, identified 1987 nominally significant CpGs, achieving AUCs of 0.73–0.76, though harmonized data reduced

predictive power, highlighting preprocessing challenges [112].

In a recent meta-analysis, researchers examined the use of whole-blood DNA methylation data to distinguish between healthy individuals and those with diseases, using PD and SZ as case studies. The study highlighted the significant role of data harmonization in enhancing classification accuracy, noting improvements of up to 20% when training and test datasets were processed using different methods. Remarkably, the best classification performance for PD was achieved using LightGBM, a gradient-boosted decision tree ensemble, which reached an accuracy of over 97% on harmonized data. In contrast, for SZ, the highest accuracy was 72%, achieved with CatBoost on harmonized data. Additionally, the study demonstrated that dimensionality reduction could effectively reduce the number of features without sacrificing classification accuracy. The use of kNN with a single neighbor was identified as the optimal imputation method, maintaining baseline accuracy levels in datasets with no missing values [96]. For very late-onset schizophrenia-like psychosis, a diagnostic XGBoost model using methylation microarray data ($n=1218$) and bisulfite sequencing, prioritized by methylation quantitative trait loci (mQTL)and linkage disequilibrium patterns, achieved perfect discrimination (AUC 1.0) from SZ and AD, driven by epigenetic alterations in the GNB5 gene [153]. Similarly, a diagnostic framework classified multiple degenerative disease such as AD, PD, and ALS with average AUC 97.4% [154]. The diagnostic framework for this study involved a multi-step process: initially, a RFmodel filters 12,578 CpG sites from Illumina 27K/450K EPIC data, followed by a multi-1D CNN model refining 600 key biomarkers from 6438 CpG sites and reduced to 110 key biomarkers with biological analysis, including GO enrichment and KEGG pathway analysis, for input into a ResNet-based multiclass prediction model [154].

In parallel, another research effort employed a sparse partial least squares discriminant analysis (SPLS-DA) model to classify SZ from controls using DNA methylation data captured on a 450k array. The SPLS-DA model, which incorporates LASSO regularization, operates on the premise that a small subset of variables primarily influences the underlying biological processes. This model achieved a positive predictive value of 80%, outperforming the classification efficacy of polygenic risk scores [155]. Similarly, research on major depression led to the development of a LASSO penalized regression model using a large EPIC dataset. This model proved comparably effective against polygenic risk scores. Further analysis explored the correlation of methylation risk scores (MRS) with 61 behavioral phenotypes, revealing that while polygenic risk scores were linked to psychosocial and mental health phenotypes, methylation risk scores showed a stronger association with lifestyle and sociodemographic factors [156]. Another AD study developed a MRS using a robust linear model with 19 CpGs and 24 DMRs from the Healthy Brain Initiative cohort (n = 88), analyzed via the comb-p method with Sidak correction, predicting cognitive decline with significant enrichment in lipid metabolism and synaptic plasticity pathways [157].

Recent studies have significantly advanced the predictive potential of DNA methylation profiling in neurological disorders by integrating cfDNA with sophisticated ML models to identify epigenetic biomarkers, though clinical translation remains limited. For AD, an ANN model trained on cfDNA methylation profiles identified 3684 differentially methylated CpGs enriched in pathways like glutamatergic synapses and axon guidance, achieving AUCs of 0.949–0.998 across six AI platforms including SVM, RF, generalized linear model, prediction analysis for microarrays, linear discriminant analysis, and ANN, with 94.5% sensitivity and specificity in an independent test group [158]. In another AD study, a logistic regression model implemented in MetaboAnalyst, analyzing cfDNA from 25 AD patients and 23 controls, identified 130 differentially methylated CpGs within cytochrome P450 genes like *CYP51A1* and *CYP2S1*, achieving AUCs of 0.928 (95% CI: 0.787–1.00) in the discovery group and 0.942 (95% CI: 0.905–0.979) in the validation group, with 100% sensitivity and 90% specificity [159]. In multiple sclerosis (MS), low-coverage WGBS of cfDNA from 75 patients was analyzed using seven ML models, with a GradientBoostingClassifier (n_estimators = 200, learning_rate = 0.2, max_depth = 10, subsample = 0.8) for MS vs. controls achieving AUCs of 0.70–0.82 using top two SVD components of DMR methylation density, and a LightGBM classifier (n_estimators = 100, learning_rate = 0.1, num_leaves = 31) with recursive feature elimination for progressive vs. relapsing–remitting MS classification, outperforming traditional biomarkers like neurofilament light chain (AUC = 0.67–0.81); a linear mixed-effects model identified 5392 prognostic regions predicting disability progression (AUC = 0.74) within three years [160]. Although these studies show promising results in the classification of neurological disorders, they have so far been limited to predictive applications and have not yet been translated into validated clinical diagnostic tests.

#### *Cardiovascular and metabolic diseases*

Recent advancements in epigenetic profiling also have enhanced the predictive power for metabolic diseases by integrating DNA methylation data with ML models, identifying novel biomarkers and pathways. In a study on gestational diabetes mellitus, whole-genome methylation

datasets from neonates born to gestational diabetes mellitus mothers were analyzed, identifying DMRs in genes like *PPARG* and *INS*, enriched in insulin signaling, *AMPK* activation, and adipocytokine signaling pathways; a RF model achieved AUCs of 0.89 in public data and 0.82 in a clinical cohort, suggesting potential for early risk assessment despite limited sample size [161]. For diabetic populations, a 20-year follow-up study from the National Health and Nutrition Examination Survey ($n=454$) employed eight ML models (logistic regression, decision tree, RF, XGBoost, AdaBoost, GBM, SVM, and multilayer perceptron), including Cox regression and restricted cubic spline analysis, to identify four DNA methylation-derived epigenetic biomarkers, with hazard ratios ranging from 2.86 to 4.08 for all-cause and cardiovascular mortality, demonstrating strong associations with mortality risk ($P<0.001$) [162].

In recent research on coronary heart disease (CHD), a study utilizing the Framingham Heart Study 450k array dataset identified five key DNA methylation-regulated genes through dimensionality reduction techniques. These genes, *ATG7, BACH2, CDKN1B, DHCR24*, and *MPO*, were instrumental in developing predictive ML models for CHD. Utilizing methods such as LightGBM, XGBoost, and RF, the study found that the LightGBM model demonstrated the most favorable predictive performance. Specifically, in the validation set, this model achieved an AUC of 0.834, with sensitivity and specificity values of 0.672 and 0.864, respectively [163]. In a related earlier study, also based on the Framingham Heart Study, researchers incorporated four DNA methylation sites, two single nucleotide polymorphisms (SNPs), and age and gender into an RF classifier. This approach yielded an accuracy of 78%, with sensitivity and specificity rates of 0.75 and 0.80, respectively, in detecting CHD [164]. In heart failure with preserved ejection fraction, the HFmeRisk deep learning framework, incorporating LASSO, extreme gradient boosting, and a factorization-machine-based neural network, combined 25 CpG loci with five clinical features from the Framingham Heart Study, achieving an AUC of 0.90 (95% CI: 0.88–0.92) and a Hosmer–Lemeshow statistic of 6.17 ($P=0.632$), outperforming models using only clinical or methylation data and revealing pathways like amino acid metabolism and intercellular signaling [111].

In recent advancements in osteoarthritis research, a study leveraged both Ridge and LASSO regression techniques to develop a parsimonious model using data from the Osteoarthritis Biomarkers Consortium (OABC) 850k array dataset. This model adeptly classified various progression types of knee osteoarthritis, including future radiographic-only progression (AUC=0.94), pain-only progression (AUC=0.97), combined pain and radiographic progression (AUC=0.79), and any form of progression (AUC=0.86), demonstrating its robust predictive capability [165]. In a distinct study that utilized vast electronic health records coupled with ML techniques such as LASSO, elastic net, and ridge regression, researchers at UCLA Health Biobank developed MRS. These scores, which represent a linear combination of methylation states, were generated for 607 phenotypes encompassing clinical lab tests, medication use, and medical diagnoses based on 831 samples. When these MRS were incorporated into a baseline set of predictive features, they significantly enhanced the prediction of 139 outcomes, markedly outperforming polygenic risk scores (PRS), which improved predictions for only 22 outcomes [166]. The rapidly evolving field of cardiovascular epigenetics urgently requires standardized protocols, as their absence impedes robust data integration and comparison, restricting the clinical implementation of DNA methylation risk scores [167].

#### *Rare diseases*

Emerging research on rare diseases leverages DNA methylation profiling and ML to uncover unique epigenetic signatures, offering novel diagnostic insights into complex disorders. In the current research on recurrent constellations of embryonic malformations (RCEM), a binary SVM classifier model with 98.32% specificity was developed based on the combined episignature of peripheral blood DNA methylation [168]. The researchers also assessed the correlation of RCEM with other disorders listed in the EpiSign Knowledge Database. By highlighting genomic regulatory pathways specifically involved in the pathophysiology of RCEM, like DMRs of *POU5F1* and *MSX1*, they provided support for the emerging theory that some associations of multiple anomalies represent a single anomaly spectrum. Similarly, EpigenCentral, a freely accessible web platform (http://epigen.ccm.sickkids.ca/), facilitates interactive DNA methylation data analysis for NDDs, allowing researchers and clinicians to classify genetic variants as pathogenic or benign and analyze differential methylation patterns through a user-friendly interface, enhancing molecular diagnostics for rare disorders [169]. Additionally, in a neurodevelopmental disorder caused by biallelic loss-of-function variants in *ZNF142*, an SVM-based ML model, trained on EPIC array methylation data from 27 affected individuals and 440 controls, identified a robust DNA methylation signature comprising 88 differentially methylated probes in regulatory regions of a limited set of genes, enabling accurate classification of variants of uncertain significance and confirming the signature's consistency across peripheral blood and fibroblasts [170].

Since sequence variations are often the cause of neurodevelopmental disorders (NDD), the correct diagnosis is essential for optimally treating patients. These diagnoses can be daunting, often with VUSs or no specific congenital disease mutation. A genetic mutation in epigenetic machinery genes results in failed gene expression during critical neurodevelopment. This provides a link between genotypes and clinical phenotypes, leading to an innovative concept of discovering syndrome-specific molecular episignatures instead of phenotypes to enhance diagnosis [171]. Each NDD has a unique episignature that represents genetic changes underlying the disease's pathogenicity, leading to a diagnosis, and identifying novel pathogenic variants. Recent progress has seen the use of algorithms for the classification of diseases with DNA methylation analysis, and a special form of this service, EpiSign, is designed for congenital neurodevelopmental disorders [172–175].

EpiSign uses a methylation microarray technique on blood-derived DNA from NDD cohorts and controls to create the EpiSign Knowledge Database (Fig. 4). It then inputs the data into its ML-based algorithms to compare the DNA methylation patterns of individuals suspected to have NDDs. EpiSign is currently in its fourth version, and it detects over 90 congenital NDD conditions and more than 96 genes or genetic regions associated with NDDs. The list will continue to grow as more methylation NDD cohorts are included in the EpiSign Knowledge Database. Due to its disease-specific epigenetic signatures, this assay offers high sensitivity and specificity, making it a robust tool for diagnosing conditions that cannot be diagnosed using traditional genetic assays. Considering how widely spread epigenetic mechanisms are across the genome, the broad utility of the assay encompasses a wide range of disorders, especially those with intricate epigenetic patterns, like imprinting syndromes and some intellectual disabilities. The assay boosts the diagnostic yield when working in combination with traditional genetics tests, including exome sequencing or genome sequencing, as it uncovers causes that are not sequence-related but in epigenetic regulations over DNA sequences. It uses noninvasive and simple blood samples, allowing data integration into usual clinical workflow, is substantially accessible in both cost and expertise, provides valuable insights into the underlying mechanisms of disease, and aids research and the development of targeted therapies [173]. The identification of specific epigenetic markers can have prognostic implications, helping predict disease progression and treatment response and, in some cases, guiding therapeutic decisions in personalized medicine. The assay is designed for efficient processing, offering a rapid turnaround time crucial for clinical decision-making [12]. Additionally, it is cost-effective compared to some high-throughput genomic technologies, often leading to a definitive diagnosis without the need for further extensive testing. These benefits collectively make the EpiSign assay a valuable asset in clinical settings for the diagnosis and study of epigenetic disorders.

Despite its promising potential, EpiSign also has some limitations that should be considered. It has only been validated on DNA from peripheral blood samples, so it may not reflect some tissue-specific episignatures, which can only be measured in the target tissue [176]. This method addresses a significant challenge in the field due to the scarcity of Mendelian disorders, particularly given the broad spectrum of rare diseases globally. Generating detection models demand cohorts with gene-specific pathogenic variants, feasible for prevalent disorders, yet occasionally leading to inconclusive results due to the mild scale of certain episignatures and the cohort size's impact on assay sensitivity [11]. Hence, there are moderate episignatures, and some newly mapped syndromes, such as Genitopatellar syndrome, Coffin-Siris syndrome-4, and Coffin-Siris syndrome-1,2, still require

(See figure on next page.)

**Fig. 4** Comprehensive workflow for epigenetic signature analysis and methylation profiling in syndromic diseases. **A** Epigenetic Mechanisms and Syndromic Variations. This panel illustrates the influence of mutations in epigenetic modifiers (writers, readers, erasers, and remodelers) on the epigenetic signatures of various syndromes (X, Y, Z). Differential epigenetic modifications across four genes (A, B, C, D) are depicted, contrasting a reference epigenetic signature with those altered by specific mutations (MUT) in the modifiers. Panel **B** DNA Methylation Analysis Process. The workflow begins with DNA extraction, followed by bisulfite treatment which converts unmethylated cytosines to uracil, depicted as C to U transition. Post-deamination, DNA undergoes amplification where the methylated cytosine remains unchanged, and the uracil is read as thymine. This is followed by hybridization to allele-specific probes (Infinium I and II) that differentiate between methylated and unmethylated loci, culminating in a high-resolution methylation profile. **C** Data Processing and Epigenetic Signature Classification. Methylation data are converted from $\beta$-values to M-values using a logarithmic transformation for enhanced analysis precision. A multivariable linear regression model is applied to identify regions with significant methylation alterations, adjusted for multiple testing. Significant features are further refined by removing highly correlated probes. The processed data are then used to train a support vector machine (SVM) for the classification of methylation profiles into known syndromic categories. Unsupervised learning (t-SNE) visually assesses the clustering of methylation signatures, cross-referenced against an epigenetic design knowledge database. **D** Syndrome Diagnosis and Detection: Methylation variant pathogenicity (MVP) scores are computed for each syndrome based on the derived methylation profiles. These scores are plotted, and the potential disorder is identified through the highest-scoring category, providing a diagnostic assessment based on epigenetic biomarkers

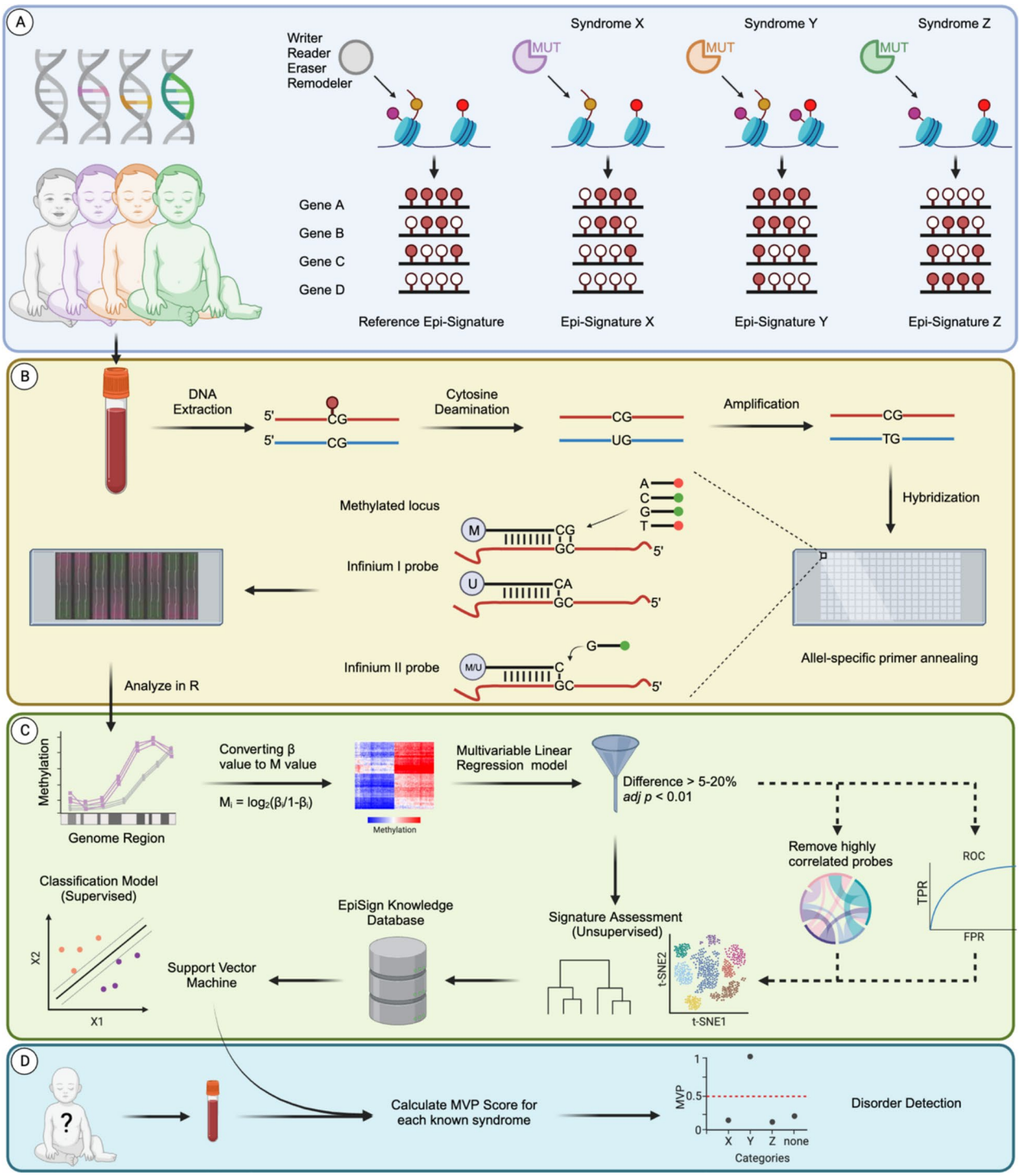


**Fig. 4** (See legend on previous page.)

further work to achieve higher accuracy. The lack of sensitivity for these conditions is due to their recent mapping and the small sample size. As more samples are collected, the accuracy is expected to improve. Additionally, the accuracy of EpiSign can be influenced by factors such as technical variability in DNA methylation analysis, sample quality, and the choice of classification algorithm [57, 177]. A recent study presented the inaugural, impartial assessment of 16 episignatures spanning 10 neurodevelopmental disorders, focusing on predictive accuracy and

consistency. The study involved DNA methylation data from 57 NDD patients and 25 healthy controls. Utilizing a leave-one-out scheme with a k-nearest-neighbor classifier and a combination of published episignature data and new validation data, the study offered unbiased estimates of specificity and sensitivity for each signature. While achieving 100% specificity, sensitivities varied widely among signatures. Notably, certain signatures like *ATRX, DNMT3A, KMT2D*, and *NSD1* demonstrated 100% sensitivity, while others like *CREBBP-RSTS* and a *CHD8* signature fell below 40%. Some signatures showed 70–100% sensitivity but exhibited instability due to methylation profile heterogeneity and rare discordant samples [178].

There are some biological challenges associated with the use of EpiSign. First, the challenge of mosaic pathogenic variants, accounting for 5–10%, complicates the clinical utility of episignatures and genetic tests. If mutations affect multiple tissues early in development, methylation differences may manifest in peripheral blood, albeit at lower levels reflective of mosaicism. Episignatures with more pronounced methylation differences might detect lower levels of mosaicism, but those impacting specific tissues, like neural tissue, may elude detection using peripheral blood samples. Determining detection thresholds for each episignature in mosaic cohorts requires further comprehensive analysis [179]. The EpiSign assay's limitation lies in its inability to readily detect low-level mosaic imprinting disorders observed in the validation cohort. Detection is feasible only at levels > 20%, which is attributed to normal control variability and poses challenges in identifying mosaicism-related discordant samples [11]. While EpiSign exhibits 100% sensitivity and specificity for detecting fragile X syndrome in male patients and differentiating patients with mosaic methylation defects, it falls short in detecting FMR1 methylation changes among full mutation and premutation carrier females [180]. Notably, in X-linked disorders such as CdLS5, a concealed episignature was observed, potentially due to skewed X-inactivation. Specifically noted in CdLS5, individuals with skewed X-inactivation of the chromosome carrying the mutated HDAC8 allele lacked discernible methylation profiles, particularly observed in peripheral blood. The study's cohort comprised solely female CdLS5 subjects, indicating the necessity for exploring potential methylation patterns in male CdLS5 subjects, yet unstudied due to the absence of X-inactivation [176].

The scarcity of data for rare genetic developmental disorders poses a significant challenge in developing robust ML models for DNA methylation-based diagnostics, often leading to overfitting due to limited cohort sizes and heterogeneous methylation profiles. In 2025, innovative solutions have addressed these issues by leveraging advanced sequencing and synthetic data approaches. Nanopore long-read WGBS enabled simultaneous detection of episignatures (NBSepi), SNPs, structural variants, and X-inactivation patterns in 20 patients with developmental disorders, with an SVM classifier achieving accurate identification in 17/19 cases with pathogenic variants, consolidating multi-step diagnostics into a single assay [181]. Additionally, the MethaDory tool introduced a train-on-synthetic-test-on-real approach, using SVM classifiers trained on 169 synthetic cases derived from anonymized DNA methylation data of unaffected individuals and median beta values from 89 disorders, improving classification accuracy for variants with intermediate effects and mosaic cases without requiring private patient data, outperforming platforms like EpigenCentral and NBSepi [182]. These advancements mitigate data scarcity and overfitting risks, enhancing the potential for scalable, privacy-compliant diagnostic models.

#### Diagnosis

In recent years, hundreds of studies on disease diagnosis and classification using ML on DNA methylation data have been documented in PubMed and Google Scholar. However, only a few models, such as Galleri and EpiSign, are currently utilized in clinical diagnostics. Following the successful implementation of ML methods on large-scale DNA methylation datasets, two general diagnostic approaches have emerged for clinical translation. The first involves selecting a limited number of methylation marker sites as features with the highest contribution to model classification, followed by using PCR-based methods for patient sample analysis. These features are classified using statistical thresholds or simple linear models. This approach is cost-effective, easily deployable in most diagnostic laboratories, and straightforward for laboratory personnel to learn. However, it is diagnostically limited to one or a few disease conditions. In contrast, the second approach employs high-throughput array and bisulfite sequencing tools to analyze blood samples and compare methylation signatures against extensive databases (Fig. 4D). Although this method is expensive, requiring samples to be sent to genomics centers and relying on highly skilled laboratory professionals, it enables the simultaneous diagnosis of numerous disease conditions from a single test.

Currently, validated diagnostic tests offered by biotechnology companies that follow the first approach (PCR-based methods) are primarily designed to detect a single type of cancer. For example, the FDA-approved Cologuard®, a multi-target stool DNA test for colorectal cancer (CRC) screening in adults aged 45 and older at average risk. Recommended every three years, the test detects methylation changes at the *BMP3* and *NDRG4*

promoters, along with seven point mutations in the *KRAS* gene. Utilizing the Quantitative Allele-specific Real-time Target and Signal Amplification (QuARTS™) technology combined with a logistic regression algorithm, Cologuard® has shown a sensitivity of 92% and a specificity of 87% for CRC detection and 42% sensitivity for advanced adenomas [183]. Similarly, Epi proColon®, another FDA-approved CRC-focused test, identifies methylation of the *SEPT9* gene in cfDNA using a proprietary PCR-based method. Though slightly less sensitive than Cologuard® (sensitivity of 68–72% and a specificity of 80–82%), it offers a simpler and less invasive screening option for populations with low compliance to traditional methods [184, 185]. In prostate cancer diagnostics, ConfirmMDx®, as a Laboratory Developed Test (LDT), analyzes the methylation status of *GSTP1*, *APC*, and *RASSF1* genes in histologically normal prostate tissue to identify cancer's field effect. This test reduces unnecessary repeat biopsies by up to 65%, maintaining a high negative predictive value of 90% [186, 187]. On the other hand, the IvyGene® Cancer Blood Test is currently marketed as a LDT. This test assesses the methylation status of *MYO1G* and *TNFAIP8L2* genes in cfDNA using an NGS approach. Initially validated with plasma from 197 subjects, either cancer-free or diagnosed with breast, colon, lung, or liver cancer, the test demonstrated a sensitivity of 84% (95% CI 75–93) and specificity of 90% (95% CI 85–95) for detecting these four cancer types. Preliminary findings also suggest that a positive test result could indicate the presence of other cancer types [185].

Galleri® is a multi-cancer early detection (MCED) test developed by GRAIL that analyzes methylation patterns in cfDNA to detect a shared cancer signal across more than 50 types of cancer, and was explored in previous section, is one of the example of second approach (high-throughput-based methods). The Circulating Cell-free Genome Atlas (CCGA) study (NCT02889978) has provided a robust foundation for validating the performance of Galleri®, a MCED test that leverages cfDNA methylation patterns for cancer detection and localization. In a post hoc analysis of CCGA's third substudy, Galleri® demonstrated promising sensitivity across screened cancers (34%), unscreened solid tumors (66%), and hematologic malignancies (55%), with > 75% sensitivity for multiple cancers lacking existing population screening options, emphasizing its potential as a complementary screening tool [188]. The study also highlighted that tumor and patient characteristics such as mitotic activity, metabolic rate, and depth of tumor invasion significantly influence circulating tumor DNA (ctDNA) fractions, suggesting Galleri® is particularly sensitive for aggressive, fast-growing tumors with high mortality, many of which currently lack effective screening [189]. In its final validation phase, Galleri® showed 99.5% specificity and a 51.5% overall sensitivity for cancer detection, with sensitivity increasing by stage from 16.8% in stage I to 90.1% in stage IV. The test also achieved 88.7% accuracy in predicting cancer signal origin (CSO), covering over 50 cancer types [13]. A recent study (PROSPERO CRD42023467901) confirmed Galleri®'s performance across three studies, reporting a sensitivity of 20.8–66.3% and specificity of 98.4–99.5%, though sensitivity was lower for early-stage cancers (I–II) compared to advanced stages (III–IV). The review highlighted limitations such as the lack of completed randomized controlled trials, limited follow-up of participants with negative results, and no reported outcomes related to mortality, potential harms, health-related quality of life, or satisfaction [190]. Additionally, Galleri® research has explored cfRNA-based biomarkers to enhance detection sensitivity, especially in cases with low ctDNA levels. These "dark channel biomarkers" correlate with tumor shedding rates and offer opportunities for tumor subtype classification and tissue-of-origin localization [191]. Finally, targeted methylation analysis of cfDNA using Galleri® confirmed its ability to detect over 50 cancer types across all stages with a specificity of 99.3% and a stage-dependent sensitivity of 67.3% for 12 pre-specified cancers (stage I-III). Tissue-of-origin localization was accurate in 93% of cases with detectable cancer signals, further validating the test's clinical utility [140]. These findings collectively underscore Galleri®'s potential as a transformative tool for early cancer detection and diagnosis across a broad spectrum of malignancies. The SUMMIT Study (NCT03934866), a prospective observational cohort, extended the validation of Galleri® to assess its use alongside low-dose computed tomography for lung cancer screening in high-risk populations. In addition to validating the MCED test, the study investigated the feasibility of resource efficient communication methods for conveying pulmonary nodule results. Among 1,900 participants, 82.8% expressed satisfaction with receiving results by postal letter, while 86.3% identified it as their preferred communication method. These findings underscore the scalability and practicality of Galleri® in population-level screening programs, particularly when paired with efficient communication strategies like written correspondence [192]. Despite these promising results, Galleri® has not yet received clearance or approval from the FDA. However, according to the GRAIL website, it has been granted Breakthrough Device Designation by the FDA, which is intended to expedite the development and review of medical devices that offer significant advantages over existing alternatives. GRAIL is actively pursuing FDA approval and has initiated the REACH study to evaluate the clinical impact of Galleri® among Medicare beneficiaries, including racial and ethnic minorities, and seniors from historically underserved communities. This

study aims to enroll approximately 50,000 participants and is being conducted under an Investigational Device Exemption approved by the FDA.

Similarly, PanSeer® is a noninvasive blood test designed for the early detection of five common cancers, including stomach, esophageal, colorectal, lung, and liver, using cfDNA methylation patterns. The assay targets 477 DMRs identified through extensive analysis of publicly available datasets such as TCGA, and utilizes semi-targeted bisulfate sequencing for patients' sample. These DMRs, spanning 657 genes and 10,613 CpG sites, were validated using cancer and healthy tissue samples to ensure they represent pan-cancer signatures. PanSeer employs a logistic regression classifier with cross-validation to differentiate between cancerous and healthy samples. In the Taizhou Longitudinal Study, PanSeer detected cancer in 95% of asymptomatic individuals up to four years before conventional diagnosis, with a specificity of 96%. The test also achieved 88% sensitivity in post-diagnosis cancer patients. These findings underscore PanSeer's potential to revolutionize early cancer detection, though further longitudinal studies are needed to validate its clinical utility [193]. PanSeer® is commercialized by Singlera Genomics as an early detection MCED test and has not yet received regulatory approval for clinical use, but according to Singlera Genomics by October 2024, they were being evaluated their early detection MCED tests through large-scale population screening, with over 20,000 participants enrolled in the Taizhou Longitudinal Study. Singlera Genomics also developed PDACatch®, a liquid biopsy test for detecting pancreatic adenocarcinoma in high-risk individuals by RRBS of cfDNA. The candidate methylation haplotype blocks used for marker selection and imputed with kNN for missing measurement values as features for training SVM model. The model, incorporating 56 markers, achieved an AUC of 0.91 in detecting pancreatic cancer [194]. Recently, this assay received Breakthrough Device Designation from the FDA.

As well as Galleri® and PanSeer®, OverC® is a MCED test which analyzes methylation of cfDNA data including TCGA and GEO, with studies suggesting a sensitivity of 70–75% and specificities of 95–99%. OverC® uses SVM to discriminate between cancer and non-cancer sample then applies multiclass logistic regression to predict primary tissue origin [195]. This test was developed by Burning Rock and received the FDA Breakthrough Devices Designation and CE In-Vitro Diagnostic. Also, OverC® is under assessment of five clinical trials until 2028 to evaluate early detection ability [196].

Another example is EPICUP®, a developed test by Grupo Ferrer Internacional SA. EPICUP® is a CE-marked diagnostic tool designed to determine the tissue of origin in patients with Cancer of Unknown Primary (CUP). Using Illumina's bead array technology, the classifier was initially developed on 2790 tumors of known origin, representing 38 tumor types and including 85 metastases, and validated on 7,691 tumor samples, achieving a specificity of 99.6% and a sensitivity of 97.7%. It demonstrated the ability to identify a primary site of cancer in 87% of CUP cases, and patients receiving site-specific therapies based on EPICUP® predictions showed significantly improved survival compared to those treated empirically. Following the discontinuation of the Infinium® Human Methylation 450K BeadChip, the newer Infinium EPIC Chip was tested and yielded identical results in studied CUP cases, ensuring the tool's continued reliability [197]. While Galleri®, PanSeer®, and OverC® use cfDNA of blood as a sample for screening and early detection, EPICUP® uses Fresh frozen or formalin fixed paraffin embedded (FFPE) tumor biopsy for tumor-type classifier [198].

One of the best example of the second approach is EpiSign®, which was described in previous section. This innovative technology leverages epigenomic data for clinical applications, marking a significant advancement in data science. As more patients utilize the service and additional data are gathered, the accuracy of classification is anticipated to be enhanced further [199]. A retrospective study of 298 patients from the Erasmus University Medical Center using the EpiSign® platform between 2019 and 2023 demonstrated its clinical utility, with targeted analyses yielding positive DNA methylation signatures in 18% of VUS, 91% of likely pathogenic variants, and 89% of pathogenic variants, while complete analyses identified disease-linked signatures in 9.0% of unsolved cases, achieving a 4.0% diagnostic yield by confirming causative variants or imprinting disorders [200]. Recently, a clinical working group with expertise in DNA methylation signature analysis on EpiSign clinical testing networks across various health jurisdictions has proposed a structured framework for this purpose, emphasizing the need for consistent reporting practices [12]. This framework categorizes findings based on the presence of an episignature and any associated gene variants, outlining specific recommendations for each scenario. 1. When an episignature is identified with high confidence and a suspected causative variant is present, it suggests a strong correlation with the episignature-associated condition. This aligns with the criteria for providing strong functional evidence of a gene's role in disease manifestation. 2. The recommendation is to pursue additional genomic testing for cases where an episignature is positively identified with high confidence but no causative variant is detected. This scenario underscores the complexity of genetic diagnostics and the potential for undiscovered or cryptic variants impacting gene

function. 3. If an episignature is detected with a moderate confidence and a causative variant is present, it is suggested that the finding could be due to variants with a partial functional impact. This highlights the variability in genetic expression and its influence on disease phenotypes. 4. When an episignature is found with moderate confidence without a causative variant, the interpretation suggests a potential diagnosis associated with the episignature, recommending further molecular testing. 5. Inconclusive episignature findings, regardless of the presence of a causative variant, point to the need for additional investigation. These findings reflect the challenges in drawing clear conclusions from methylation profiles that only partially match known episignatures. 6. A negative episignature finding, whether or not a causative variant is present, indicates that the methylation profile does not match the known episignatures for the condition being tested. However, this does not rule out the clinical diagnosis due to the potential for multiple episignature profiles associated with various conditions.

#### Treatment

Identifying the tissue of origin in cancers, especially metastatic or CUP, is critical because the origin directly influences therapeutic strategies. Different tissues exhibit unique molecular and biological characteristics, which dictate distinct responses to targeted therapies, chemotherapy, and immunotherapy. Accurate identification of the tissue of origin enables oncologists to tailor treatment regimens that optimize efficacy and minimize adverse effects. For example, metastatic colorectal cancer (mCRC) originating from the gastrointestinal tract often expresses the epidermal growth factor receptor (EGFR). Anti-EGFR therapies like cetuximab or panitumumab are effective treatments for mCRC when RAS mutations are absent. In contrast, if the metastases are derived from non-small-cell lung cancer (NSCLC), the therapeutic approach would shift toward using tyrosine kinase inhibitors (e.g., osimertinib) for EGFR-mutant cases or immune checkpoint inhibitors like pembrolizumab for PD-L1-expressing tumors [201, 202]. Similarly, breast cancer metastases often retain hormone receptor (HR) and HER2 expression patterns reflective of the primary tumor. Identifying these markers allows for hormone therapies such as tamoxifen or aromatase inhibitors in HR-positive cases, or HER2-targeted therapies like trastuzumab in HER2-positive cases. However, treating a metastasis from ovarian cancer would typically involve platinum-based chemotherapy due to the different biological behavior of ovarian malignancies [203]. Ultimately, determining the tissue of origin enables precision oncology, where therapies are tailored not just to the tumor's molecular profile but also its biological context. As more advanced techniques like methylation profiling, next-generation sequencing, and proteomics are integrated into clinical workflows, they promise to enhance diagnostic accuracy and therapeutic outcomes in CUP and metastatic cancer patients. For instance, Studies have demonstrated that patients receiving site-specific therapies based on EPICUP's predictions show improved survival compared to those treated empirically [197].

While most rare diseases currently lack curative treatments, advancements in genetic testing and precision medicine have enabled interventions that can significantly improve the quality of life for affected individuals. For many conditions, the primary focus is on symptom management, supportive care, and lifestyle modifications tailored to the specific needs of patients. Furthermore, genetic insights play a critical role in prenatal diagnosis and decision-making for at-risk families, guiding reproductive planning and early interventions. For example, Prader-Willi syndrome, caused by the loss of function of paternal genes on chromosome 15q11-q13, is characterized by intellectual disabilities, hyperphagia, and obesity. While there is no cure, early interventions such as growth hormone therapy, tailored behavioral management, and using sugar-free diet can significantly improve outcomes. Genetic testing enables early diagnosis and facilitates carrier screening and counselling for at-risk relatives, especially for planning subsequent pregnancies [204]. Similarly, Angelman syndrome, resulting from maternal deletions or mutations in the *UBE3A* gene, presents with severe developmental delays and a unique behavioral phenotype. Supportive therapies, including speech and physical therapy, alongside targeted educational interventions, can enhance the quality of life for these individuals. Prenatal diagnosis through genetic testing provides families with critical insights for decision-making [205]. Another example is Coffin-Siris syndrome, a rare genetic disorder associated with mutations in the *ARID1B* or related genes, leading to developmental delays, intellectual disabilities, and characteristic facial features. Genetic testing is essential for confirming the diagnosis and providing families with a clear understanding of inheritance patterns. This knowledge can guide at-risk families in exploring reproductive options such as pre-implantation genetic testing to prevent recurrence in future pregnancies [206]. These examples highlight the growing importance of genetic testing and counselling in rare diseases, not only for improving patient outcomes but also for empowering at-risk families with critical information for future planning. As technologies like DNA methylation profiling and machine learning-based approaches, such as the EpiSign assay, become more integrated into

clinical practice, they promise to enhance our understanding of rare diseases. These tools enable earlier and more precise diagnoses by leveraging disease-specific epigenetic signatures, offering hope for tailored interventions and better-informed reproductive planning for at-risk families.

#### Follow-up clinical parameters

The ability to analyze molecular information from tumors through liquid biopsy has revolutionized the monitoring of MRD following curative therapies. Liquid biopsy enables sequential sampling, providing a noninvasive method for tracking dynamic changes in DNA methylation patterns, which are indicative of residual disease or early relapse. For instance, in non-muscle-invasive bladder cancer, where recurrence rates can reach 50–70% within five years, the use of urine-based assays has simplified lifelong surveillance. The Bladder EpiCheck® test (Nucleix Ltd.), a CE-certified assay based on 15 proprietary DNA methylation biomarkers, has demonstrated clinical utility in the cancer follow-up. Studies have shown its sensitivity to be 68.2% (95% CI, 52.4–81.4%) and as high as 91.7% (95% CI, 73.0–99.0%) when low-grade noninvasive low-grade papillary bladder tumors are excluded, with a specificity of 88% (95% CI, 83.9–91.4%) [207–209]. This reduces reliance on invasive cystoscopy and facilitates more efficient postoperative surveillance. Additionally, in lung cancer, a study using the tumor-informed methylation-based MRD model (timMRD) demonstrated superior accuracy in detecting recurrence and predicting disease-free survival. Analysis of ctDNA in the MEDAL cohort ($n=195$) and validation in the DYNAMIC cohort showed timMRD scores strongly correlated with tumor burden and somatic mutations. Patients with elevated postoperative timMRD scores had significantly shorter disease-free survival (HR = 3.08, $P=0.002$). Compared to somatic mutation-based ctDNA analysis, timMRD was more effective in identifying relapse, particularly in stage I and baseline ctDNA-negative patients, achieving a 97.2% negative predictive value up to 120 days prior to relapse [210]. In pediatric acute lymphoblastic leukemia (ALL), DNA methylation profiling combined with supervised machine learning, specifically random survival forests, has been used to predict relapse and mortality risk. A relapse risk predictor based on 16 CpG sites and a mortality risk predictor using 53 CpG sites were developed from a cohort of 763 patients, achieving c-indexes of 0.667 and 0.751, respectively. Validation in independent Canadian and Nordic cohorts confirmed the models' prognostic value, with the integration of clinical risk group data further enhancing predictive accuracy. These findings highlight the potential of DNA methylation as a robust tool for refining risk stratification and enabling personalized treatment strategies in pediatric ALL [32]. Furthermore, in AML, the integration of machine learning with DNA methylation data has shown promise in improving MRD detection. By analyzing methylation patterns, it is possible to identify residual leukemic cells that may not be detectable through conventional methods, thus providing a more accurate assessment of remission status [211]. Collectively, these approaches demonstrate the growing role of DNA methylation biomarkers, combined with machine learning models, in enhancing MRD detection across cancer types.

Furthermore, MCED liquid biopsy tools, such as those based on circulating cfDNA, provide a minimally invasive approach for longitudinal tracking of cancer patients. For instance, methylation-based MCED tests like Galleri® not only aid in early detection, but also hold promise for follow-up monitoring of patients with a prior positive result. However, effective protocols for follow-up care remain a significant challenge [212]. A recent study employed Bayesian hierarchical models to explore sharing sensitivity data across cancer types and stages for the Galleri® test, finding strongest support for sharing sensitivity across stage IV cancers and, to a lesser extent, across cancer types within each stage when excluding low-sensitivity cancers, which improved precision of early-stage sensitivity estimates. However, high heterogeneity in ctDNA expression limited precision gains. The study also noted significant variability in Galleri®'s sensitivity (20.8% in PATHFINDER, 51.5% in CCGA3, 66.3% in SYMPLIFY) with consistently high specificity (98.4–99.5%), though evidence quality was limited by small sample sizes for many cancer types, lack of high-quality screening population data, and study design differences [213]. One critical consideration is establishing standardized pathways for clinical response after a positive MCED result. Unlike single-cancer tests with streamlined processes such as transitioning from fecal occult blood test or fecal immunochemical test to colonoscopy for colorectal cancer, MCED-positive results depend on identifying the tissue of origin, which may not always align with a specific diagnostic procedure. Therefore, regular assessment intervals must be defined for individuals with initially positive MCED results but no immediate cancer diagnosis, as well as for those with negative tests requiring periodic re-evaluation. For example, the ongoing NHS-Galleri trial, a large pragmatic study enrolling 140,000 asymptomatic individuals in England, seeks to evaluate the efficacy and follow-up of blood-based MCED tests like Galleri® in real-world settings [214]. This study is expected to provide insights into optimal follow-up care, resource allocation, and

testing intervals to ensure the utility of MCEDs in routine clinical practice. Continued collaboration among regulators, healthcare providers, and policymakers will be essential to address these challenges and maximize the clinical benefits of epigenomics-based follow-up strategies.

## Conclusion and future directions

In this review, we have demonstrated how integrating advanced DNA methylation detection techniques with ML algorithms significantly enhances our capabilities in disease detection, understanding pathogenesis, and advancing precision medicine. Starting with the fundamental concepts of DNA methylation and the various laboratory methods used for its detection, we provided a comprehensive overview of the ML process, from initial problem definition to final clinical application. Additionally, we reviewed recent studies and highlighted multiple commercially available techniques applicable to a range of diseases, including cancer, neurodevelopmental disorders, and multifactorial diseases such as Alzheimer's and cardiovascular diseases.

As previously mentioned, the nature of the research question will guide the selection of the appropriate model and architecture for DNA methylation detection and interpretation. While classic ML models like SVM and RF have demonstrated significant clinical outcomes, their limitations are well-documented. On the other hand, more complex ML architectures, specifically deep learning models such as transformers and autoencoders, are being increasingly utilized to tackle more sophisticated questions [61, 151, 215–218]. It is crucial to emphasize that the choice of neural networks should be driven by their suitability for the data and the research question at hand rather than a mere trend-following approach [219]. The model selection process is both technical and strategic, necessitating a deep understanding of the research aims, whether it's identifying disease biomarkers, predicting clinical outcomes, or analyzing gene expression mechanisms. The type of model chosen is inherently tied to the research question, such as searching for associations between methylation patterns and disease status, predicting clinical outcomes based on methylation profiles, or understanding the regulatory cascade of gene expression.

As highlighted in Sects. "Diagnosis" and "Follow-up clinical parameters", all clinically deployed DNA methylation-based diagnostic models currently rely on traditional ML methods, while DL approaches remain confined to research settings. Despite the notable successes of DL in other fields, its performance in DNA methylation-based medical diagnostics has been comparatively limited due to several fundamental challenges inherent to the nature of methylation data, the complexity of DL architectures, and the requirements for clinical applicability. Firstly, DNA methylation datasets are characterized by a high dimensionality-to-sample ratio, often containing hundreds of thousands of CpG sites with limited sample sizes. Deep learning algorithms require vast and diverse datasets for optimal performance, whereas the sparse nature of methylation data increases the risk of overfitting. In contrast, traditional ML models, such as RF and SVM, are better suited for small-sample, high-dimensional data, as they leverage effective regularization techniques and feature selection capabilities. Secondly, class imbalance is a recurring issue in DNA methylation datasets, where disease-specific CpG sites may be underrepresented. DL models generally underperform in handling imbalanced datasets compared to traditional methods like Gradient Boosting and SVM, which have proven robust in such scenarios. Moreover, the sparsity of informative CpG sites further complicates DL approaches. Traditional ML models, such as LASSO regression or feature-selection-optimized SVMs, can identify and prioritize a small subset of CpG sites that contribute the most to disease classification, a task at which DL often struggles due to its reliance on capturing complex, global patterns. Another critical factor is the interpretability of results. In medical diagnostics, transparency and result explainability are paramount for clinical adoption. Models like logistic regression and RF allow for clear identification of key CpG sites and their respective contributions to predictions, facilitating trust and actionable insights for clinicians. Conversely, DL models typically function as "black boxes," offering little transparency, which limits their clinical utility. The interpretability of traditional ML algorithms thus makes them more suitable for DNA methylation data, where feature importance plays a central role in clinical decision-making. Future research must address these limitations to unlock the full potential of DL in DNA methylation-based diagnostics. Efforts to generate larger, more balanced, and diverse datasets through multicenter collaborations and integration of data repositories will help overcome DL's dependency on extensive training data. Hybrid models that combine the feature-selection strengths of traditional ML with the pattern-recognition power of DL may offer a promising direction. Additionally, the development of explainable AI techniques tailored for DL models could improve transparency and foster clinical adoption.

The decreasing costs of sequencing technologies and the integration of high-throughput genomics have paved the way for the generation of large-scale multi-omics datasets. This advancement positions DL methodologies as superior to traditional ML approaches, particularly when handling the vast and intricate nature of such data.

DL's capacity to model complex, nonlinear relationships renders it especially effective in this context [220]. Moreover, the precision of these technologies has advanced to the single-cell level, enabling spatial molecular analyses. Although current research has demonstrated the potential of DL in this domain, challenges persist. Complex protocols, high costs, significant technical and biological noise, and limited reproducibility have hindered the widespread adoption of multi-omics approaches in clinical laboratories, confining them largely to hypothesis-generating research. Nonetheless, with ongoing technological advancements, anticipated reductions in costs, and improvements in user-friendliness and accuracy, the future holds promise. These developments are expected to facilitate the incorporation of DL models into DNA methylation-based diagnostic tools, thereby enhancing clinical decision-making and patient care [221, 222]. Nonetheless, with ongoing technological advancements, anticipated reductions in costs, and improvements in user-friendliness and accuracy, the future holds promise. These developments are expected to facilitate the incorporation of DL models into DNA methylation-based diagnostic tools, thereby enhancing clinical decision-making and patient care [223].

Similarly, the specific preference observed in utilizing biochemical technique for detection DNA methylation pattern in clinical tests. For instance, while MeDIP-seq offers a broader coverage of the methylome, it lacks the single-base resolution provided by WGBS and the targeted precision of Infinium arrays. This limitation can result in less accurate detection of differentially methylated regions, particularly in CpG-dense areas [224]. In clinical diagnostics, precision and reproducibility are paramount. Infinium MethylationEPIC arrays and WGBS deliver high-resolution, quantitative data essential for identifying specific epigenetic biomarkers associated with disease states. MeDIP's lower resolution and potential biases in CpG-rich regions may compromise its diagnostic accuracy, making it less suitable for clinical applications where exact methylation mapping is critical [225]. Additionally, the reproducibility of MeDIP can be affected by variations in antibody affinity and experimental conditions, leading to inconsistencies in methylation profiles. In contrast, Infinium arrays and WGBS have well-established protocols and quality control measures that enhance their reliability in clinical settings [226].

Looking ahead, we see two potential paths for future developments. The first path continues to explore and refine classic ML techniques, while the second path delves deeper into innovative deep learning models. The ongoing generation of more extensive datasets will benefit both paths, though deep learning models, in particular, stand to gain significantly as they are currently constrained by sample size limitations and the scarcity of cases, especially for rare diseases. In this context, we have thoroughly discussed the ML process and strategies for improving model accuracy and achieving favorable clinical outcomes. Here, we will focus on the future perspectives and limitations of deep learning models. One significant challenge is the black box nature of these models, which complicates their clinical applicability. Future work in interpretable/explainable AI, exemplified by approaches like mEthAE, XAI-AGE, and CFA, holds promise in addressing this issue [227–229]. Many deep learning architectures, particularly large language models currently used for genome interpretation, were not originally designed for biological tasks. However, their performance in this domain has been remarkable. An exciting future direction lies in designing algorithms that are initially customized for the unique challenges and nuances of biological data and inquiries. It is also essential to recognize that ML is not a one-size-fits-all solution. It is crucial to discern when it is not appropriate to use these methods, such as when data is insufficient, when the goal is to gain understanding rather than make predictions, or when performance assessment cannot be done fairly [219, 230].

Two paths can be framed as Track 1 (interpretable classical ML) and Track 2 (advanced deep learning). In track 1, robust, transparent models will keep delivering compact CpG panels with external validation and simpler lifecycle management. In multi-cancer early detection and related blood-based tests, simple algorithms such as logistic regression and tree-based methods perform competitively and remain clinically readable [231], and new cross-platform frameworks show that explainable, lightweight classifiers can match or exceed heavier models while easing deployment across assay types [110]. Priorities include leakage-proof feature selection, batch harmonization, and prospective multi-site studies aligned to regulatory expectations. In track 2, transformer foundation models pretrained on large methylome compendia now learn context-aware CpG embeddings and transfer to imputation, aging, tissue-of-origin, and disease-risk tasks with limited fine-tuning [15, 16]. These models are well-positioned for small-cohort clinical problems and for multi-omics fusion with accessibility, histone marks, and expression. Architectures that encode biology upfront, such as pathway- or region-aware networks, will improve both performance and credibility. In practice, linear models probably remain well-suited for rare diseases with limited samples, deep learning excels in multifactorial and multi-omic contexts, and low-stage cancer detection is likely to advance through improvements in sequencing technologies. We hope that future models support multi-disease and multi-panel detection in a single assay.

Trust and transparency are non-negotiable. Black-box models must ship with explanation layers that attribute predictions to CpGs, regions, and pathways, plus clear model cards, data sheets, and drift monitors. Interpretable overlays for the brain-tumor methylation classifier show that clinically useful attribution is feasible [10, 21]. SHAP remains a practical standard for feature-level explanations in regulated settings. Agencies now emphasize good machine learning practice and transparency for AI-enabled medical devices, which pipelines will need to address explicitly [232].

Agentic AI above both tracks. LLM-based agents are starting to orchestrate end-to-end omics workflows: planning analyses, invoking tools, enforcing provenance, and drafting clinician-facing reports under human oversight. Research agents can already design and iterate on genetic-perturbation experiments and outperform classic optimization baselines in closed-loop discovery. Early clinical prototypes of autonomous decision-support agents in oncology have been validated prospectively, but deployment requires guardrails: tool-use allow-lists, sandboxed execution, versioned workflows, pre-registered analysis plans, and human sign-off [233]. The same "self-driving lab" paradigm is maturing and will increasingly connect agents to automated experimentation for epigenetic hypothesis testing [234]. ML should not be applied when datasets are too small or biased for fair assessment, when mechanistic insight is the primary goal, or when independent validation is impossible. In those cases, targeted experiments or simpler statistics remain safer and more informative.

Looking ahead, two promising directions hold substantial potential for advancing DNA methylation-based diagnostics. First, the development of diagnostic kits leveraging cfDNA offers the prospect of simultaneously detecting multiple diseases using a single test. Recent advancements in third-generation sequencing technologies, such as nanopore sequencing, have paved the way for real-time, high-throughput methylation analysis. However, these methods require improvements in performance, cost reduction, and reproducibility to achieve clinical reliability. Integrating multiple layers of ML algorithms or ensemble models that combine predictions from various ML approaches could enhance diagnostic precision by generating consensus probability scores. Such advancements could address current limitations in model generalizability and improve robustness across diverse populations.

Second, the adoption of nonlinear, multi-omics approaches that integrate methylation data with complementary datasets such as imaging, electrophysiological records, and other omics data (e.g., transcriptomics, proteomics) shows great promise for improving diagnostic accuracy, particularly for multifactorial diseases like cancer. A key challenge in this domain is the detection of early-stage or primary cancers, where current ML models often underperform. Feedback-driven algorithms and Bayesian methods, which iteratively update models based on low-stage sample data, could significantly enhance early detection. Moreover, advances in capturing low-concentration cfDNA from noninvasive samples (e.g., stool, urine, saliva) could further facilitate early cancer diagnosis.

Despite these promising directions, significant research gaps remain. One major challenge is the discrepancy between laboratory performance (high true positive/negative rates) and real-world outcomes, often due to population heterogeneity and limited sample sizes, particularly in rare genetic diseases [51]. Increasing sample sizes and incorporating diversity during model training and clinical trials are critical but not always feasible. Emerging approaches, such as data simulation and cross-study data banks, offer potential solutions to augment sample sizes. Additionally, many methylation sites critical for disease diagnosis are highly correlated with confounding factors, complicating model interpretation. Novel multi-omics frameworks that account for these confounders as penalty terms in ML models could mitigate this issue, but further development is needed to balance diagnostic relevance with confounder effects.

Regulatory organizations like the FDA and CE must establish more rigorous protocols to validate these ML-based diagnostic tools, ensuring their reliability in diverse clinical settings. Future research should prioritize addressing these gaps through interdisciplinary collaborations, integrating advanced sequencing technologies, and developing robust, interpretable ML models. By tackling these challenges, the field can move closer to realizing the full potential of DNA methylation-based diagnostics in precision medicine.

**Abbreviations**

| | |
|---|---|
| 5caC | 5-Carboxylcytosine |
| 5fC | 5-Formylcytosine |
| 5hmC | 5-Hydroxymethylcytosine |
| 5mC | 5-Methylcytosine |
| AD | Alzheimer's disease |
| AI: | Artificial intelligence |
| ALL | Acute lymphoblastic leukemia |
| ALS | Amyotrophic lateral sclerosis |
| AML | Acute myeloid leukemia |
| AML | Acute myeloid leukemia |
| ANNs | Artificial neural networks |
| AUC-ROC | Area under the receiver operating characteristic curve |
| CapsNets | Capsule neural networks |
| CCGA | Circulating cell-free genome atlas |
| CE | Conformité européenne |
| cfDNA | Circulating cell-free DNA |
| CHD | Coronary heart disease |
| CNNs | Convolutional neural networks |
| CPUs | Central processing units |
| CSO | Cancer signal origin |
| ctDNA | Circulating tumor DNA |

| | |
|---|---|
| CUP | Cancer of unknown primary |
| DL | Deep learning |
| DMRs | Differentially methylated regions |
| DNMTs | DNA methyltransferases |
| DWT | Discrete Wavelet Transform |
| EGFR | Epidermal growth factor receptor |
| ELSA-seq | Enhanced linear splint adapter sequencing |
| FDA | Food and drug administration |
| GDD-ENS | Genome-derived-diagnosis ensemble |
| GEO | Gene expression omnibus |
| GPUs | Graphics processing units |
| Grad-CAM | Gradient-weighted class activation mapping |
| GRU | Gated recurrent unit |
| HR | Hormone receptor |
| kNN | K-nearest neighbors |
| LDT | Laboratory developed test |
| LSTM | Long short-term memory |
| MCED | Multi-cancer early detection |
| mCRC | Metastatic colorectal cancer |
| ME | Misclassification error |
| MeDIP | Methylated DNA immunoprecipitation |
| MI | Mutual Information |
| ML | Machine learning |
| MLR | Multivariable linear regression |
| mQTL | Methylation quantitative trait loci |
| MRD | Minimal residual disease |
| MRS | Methylation risk scores |
| MS | Multiple sclerosis |
| NBSepi | Nanopore long-read WGBS enabled simultaneous detection of episignatures |
| NDD | Neurodevelopmental disorders |
| NODE | Neural oblivious decision ensembles |
| NSCLC | Non-small-cell lung cancer |
| OABC | Osteoarthritis biomarkers consortium |
| PCA | Principal component analysis |
| PD | Parkinson's disease |
| RCEM | Recurrent constellations of embryonic malformations |
| RF | Random forests |
| RNNs | Recurrent neural networks |
| RRBS | Reduced representation bisulfite sequencing |
| SAM | S-adenosyl methionine |
| scBS-seq | Single-cell bisulfite sequencing |
| scRRBS | Single-cell reduced representation bisulfite sequencing |
| SHAP | Shapley additive explanations |
| SIMPLE-seq | Simultaneous sc-5mC and sc-5hmC profiling via bisulfite-free chemical labeling |
| SMRT | Single-molecule real-time |
| SNPs | Single nucleotide polymorphisms |
| SPLS-DA | Sparse partial least squares discriminant analysis |
| SVA | Surrogate variable analysis |
| SVM | Support vector machines |
| SZ | Schizophrenia |
| t-SNE | T-distributed stochastic neighbor embedding |
| TAPS | TET-assisted pyridine borane sequencing |
| TCGA | The cancer genome atlas |
| TET | Ten-eleven translocation |
| timMRD | Tumor-informed methylation-based MRD model |
| TPUs | Tensor processing units |
| UMAP | Uniform manifold approximation and projection |
| VAE | Variational autoencoder |
| WGBS | Whole-genome bisulfite sequencing |

**Acknowledgements**
Not applicable.

**Author contributions**
EAE and AHS were involved in conceptualization and writing and editing; ABA, MEF, AH, and FG were involved in writing and prepared figures; LY and HV were involved in writing and editing; TMK, XZ, MA, and HH were involved in editing. All authors reviewed the manuscript.

**Funding**
No funding was received for this work.

**Data availability**
No datasets were generated or analysed during the current study.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interest**
The authors declare no competing interests.

Received: 4 June 2025 Accepted: 28 August 2025
Published online: 10 October 2025

**References**

1. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. Carcinogenesis. 2009;31:27–36.
2. Carter B, Zhao K. The epigenetic basis of cellular heterogeneity. Nat Rev Genet. 2021;22:235–50.
3. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet. 2013;14:204–20.
4. Wu SC, Zhang Y. Active DNA demethylation: many roads lead to Rome. Nat Rev Mol Cell Biol. 2010;11:607–20.
5. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nat Rev Genet. 2002;3:415–28.
6. He X, Zhao K, Chu X. AutoML: a survey of the state-of-the-art. Knowl-Based Syst. 2021;212:106622.
7. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8.
8. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44–56.
9. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med. 2016;375:1216–9.
10. Capper D, Jones DT, Sill M, Hovestadt V, Schrimpf D, Sturm D, et al. DNA methylation-based classification of central nervous system tumours. Nature. 2018;555:469–74.
11. Sadikovic B, Levy MA, Kerkhof J, Aref-Eshghi E, Schenkel L, Stuart A, et al. Clinical epigenomics: genome-wide DNA methylation analysis for the diagnosis of Mendelian disorders. Genet Med. 2021;23:1065–74.
12. Kerkhof J, Rastin C, Levy MA, Relator R, McConkey H, Demain L, et al. Diagnostic utility and reporting recommendations for clinical DNA methylation episignature testing in genetically undiagnosed rare diseases. Genet Med. 2024. https://doi.org/10.1016/j.gim.2024.101075.
13. Klein EA, Richards D, Cohn A, Tummala M, Lapham R, Cosgrove D, et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. Ann Oncol. 2021;32:1167–77.
14. Pyzocha NJ. Galleri test for the detection of cancer. Am Fam Phys. 2022;106:459–60.
15. Ying K, Song J, Cui H, Zhang Y, Li S, Chen X, et al. MethylGPT: a foundation model for the DNA methylome. bioRxiv. 2024;
16. de Lima Camillo LP, Sehgal R, Armstrong J, Miller HE, Higgins-Chen AT, Horvath S, et al. CpGPT: a foundation model for DNA methylation. bioRxiv. 2024;2024–10.
17. Gao S, Fang A, Huang Y, Giunchiglia V, Noori A, Schwarz JR, et al. Empowering biomedical discovery with AI agents. Cell. 2024;187:6125–51.

18. Ayub U, Naqvi SAA, Jajja SA, Afzal MU, Yum J-EI, Khakwani KZR, et al. A large language model (LLM)-based multi-agent framework for risk stratification and treatment recommendations in localized prostate cancer (locPCa). J Clin Oncol. 2025. https://doi.org/10.1200/JCO.2025.43.16_suppl.5108.
19. Ferber D, El Nahhas OS, Wölflein G, Wiest IC, Clusmann J, Leßmann M-E, et al. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. Nat Cancer. 2025;1–13.
20. Zhou J, Zhang B, Li G, Chen X, Li H, Xu X, et al. An AI agent for fully automated multi-omic analyses. Adv Sci. 2024;11:2407094.
21. Benfatto S, Sill M, Jones DT, Pfister SM, Sahm F, von Deimling A, et al. Explainable artificial intelligence of DNA methylation-based brain tumor diagnostics. Nat Commun. 2025;16:1787.
22. Parreno V, Loubiere V, Schuettengruber B, Fritsch L, Rawal CC, Erokhin M, et al. Transient loss of Polycomb components induces an epigenetic cancer fate. Nature. 2024. https://doi.org/10.1038/s41586-024-07328-w.
23. Sarhadi VK, Armengol G. Molecular biomarkers in cancer. Biomolecules. 2022;12:1021.
24. Kernaleguen M, Daviaud C, Shen Y, Bonnet E, Renault V, Deleuze J-F, et al. Whole-genome bisulfite sequencing for the analysis of genome-wide DNA methylation and hydroxymethylation patterns at single-nucleotide resolution. Epigenome Ed Methods Protoc. 2018;311–49.
25. Olkhov-Mitsel E, Bapat B. Strategies for discovery and validation of methylated and hydroxymethylated DNA biomarkers. Cancer Med. 2012;1:237–60.
26. Olova N, Krueger F, Andrews S, Oxley D, Berrens RV, Branco MR, et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol. 2018;19:1–19.
27. Cao Y, Bai Y, Yuan T, Song L, Fan Y, Ren L, et al. Single-cell bisulfite-free 5mC and 5hmC sequencing with high sensitivity and scalability. Proc Natl Acad Sci. 2023;120:e2310367120.
28. Gong T, Borgard H, Zhang Z, Chen S, Gao Z, ... Analysis and performance assessment of the whole genome bisulfite sequencing data workflow: currently available tools and a practical guide to advance DNA … Small … [Internet]. 2022; Available from: https://onlinelibrary.wiley.com/doi/abs/https://doi.org/10.1002/smtd.202101251
29. Lisanti S, von Zglinicki T, Mathers JC. Standardization and quality controls for the methylated DNA immunoprecipitation technique. Epigenetics. 2012;7:615–25.
30. Ahn J, Heo S, Lee J, Bang D. Introduction to single-cell DNA methylation profiling methods. Biomolecules [Internet]. 2021; Available from: https://www.mdpi.com/2218-273X/11/7/1013
31. Wright M, Dozmorov M, Wolen A, ... Establishing an analytic pipeline for genome-wide DNA methylation. Clin … [Internet]. 2016; Available from: https://link.springer.com/article/https://doi.org/10.1186/s13148-016-0212-7
32. Liang N, Li B, Jia Z, Wang C, Wu P, Zheng T, et al. Ultrasensitive detection of circulating tumour DNA via deep methylation sequencing aided by machine learning. Nat Biomed Eng. 2021;5:586–99.
33. Martisova A, Holcakova J, Izadi N, Sebuyoya R, Hrstka R, Bartosik M. DNA methylation in solid tumors: functions and methods of detection. Int J Mol Sci. 2021;22:4247.
34. Conlin LK, Aref-Eshghi E, McEldrew DA, Luo M, Rajagopalan R. Long-read sequencing for molecular diagnostics in constitutional genetic disorders. Hum Mutat. 2022;43:1531–44.
35. Sigurpalsdottir BD, Stefansson OA, Holley G, Beyter D, Zink F, Hardarson MÞ, et al. A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes. Genome Biol. 2024;25:69.
36. Chera A, Stancu-Cretu M, Zabet NR, Bucur O. Shedding light on DNA methylation and its clinical implications: the impact of long-read-based nanopore technology. Epigenetics Chromatin. 2024;17:39.
37. Lee I, Razaghi R, Gilpatrick T, Molnar M, Gershman A, Sadowski N, et al. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. Nat Methods. 2020;17:1191–9.
38. Sahoo K, Lingasamy P, Khatun M, Sudhakaran SL, Salumets A, Sundararajan V, et al. Artificial intelligence in cancer epigenomics: a review on advances in pan-cancer detection and precision medicine. Epigenetics Chromatin. 2025;18:35.
39. Mulqueen RM, Pokholok D, Norberg SJ, Torkenczy KA, Fields AJ, Sun D, et al. Highly scalable generation of DNA methylation profiles in single cells. Nat Biotechnol. 2018;36:428–31.
40. Chatterton Z, Lamichhane P, Ahmadi Rastegar D, Fitzpatrick L, Lebhar H, Marquis C, et al. Single-cell DNA methylation sequencing by combinatorial indexing and enzymatic DNA methylation conversion. Cell Biosci. 2023;13:2.
41. Bai D, Zhang X, Xiang H, Guo Z, Zhu C, Yi C. Simultaneous single-cell analysis of 5mC and 5hmC with SIMPLE-seq. Nat Biotechnol. 2025;43:85–96.
42. Spix NJ, Habib WA, Zhang Z, Eugster E, Milliron H, Sokol D, et al. High-coverage allele-resolved single-cell DNA methylation profiling reveals cell lineage, X-inactivation state, and replication dynamics. Nat Commun. 2025;16:6273.
43. Nichols RV, Rylaarsdam LE, O'Connell BL, Shipony Z, Iremadze N, Acharya SN, et al. Atlas-scale single-cell DNA methylation profiling with sciMETv3. Cell Genomics. 2025;5.
44. Zhu T, Liu J, Beck S, Pan S, Capper D, Lechner M, et al. A pan-tissue DNA methylation atlas enables in silico decomposition of human tissue methylomes at cell-type resolution. Nat Methods. 2022;19:296–306.
45. Zhou J, Weinberger DR, Han S. Deep learning predicts DNA methylation regulatory variants in specific brain cell types and enhances fine mapping for brain disorders. Sci Adv. 2025;11:eadn1870.
46. Zhou J, Luo C, Liu H, Heffel MG, Straub RE, Kleinman JE, et al. Deep learning imputes DNA methylation states in single cells and enhances the detection of epigenetic alterations in schizophrenia. Cell Genomics. 2025;5.
47. Christofidou P, Bell CG. The predictive power of profiling the DNA methylome in human health and disease. Epigenomics. 2025;17:599–610.
48. Siegmund KD. Statistical approaches for the analysis of DNA methylation microarray data. Hum Genet. 2011;129:585–95.
49. Warnecke PM, Stirzaker C, Melki JR, Millar DS, Paul CL, Clark SJ. Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. Nucleic Acids Res. 1997;25:4422–6.
50. Moskalev EA, Zavgorodnij MG, Majorova SP, Vorobjev IA, Jandaghi P, Bure IV, et al. Correction of PCR-bias in quantitative DNA methylation studies by means of cubic polynomial regression. Nucleic Acids Res. 2011;39:e77–e77.
51. Yang Y, Wen X, Wang L. Advancements in DNA methylation technologies and their application in cancer diagnosis. Epigenetics. 2025;20:2539995.
52. Sahoo K, Sundararajan V. Methods in DNA methylation array dataset analysis: A review. Comput Struct Biotechnol … [Internet]. 2024; Available from: https://www.sciencedirect.com/science/article/pii/S2001037024001624
53. Junaid N, Khan N, Ahmed N, Abbasi MS, Das G, Maqsood A, et al. Development, application, and performance of artificial intelligence in cephalometric landmark identification and diagnosis: a systematic review. Healthcare. 2022. https://doi.org/10.3390/healthcare10122454.
54. Kersting K. Machine learning and artificial intelligence: two fellow travelers on the quest for intelligent behavior in machines. Front Big Data. 2018;1:6.
55. Yu XT, Wang L, Zeng T. Revisit of machine learning supported biological and biomedical studies. Comput Syst Biol Methods Protoc. 2018. p. 183–204.
56. Vietri MT, D'elia G, Benincasa G, Ferraro G, Caliendo G, Nicoletti GF, et al. DNA methylation and breast cancer: a way forward. Int J Oncol. 2021;59:1–12.
57. Haghshenas S, Bhai P, Aref-Eshghi E, Sadikovic B. Diagnostic utility of genome-wide DNA methylation analysis in mendelian neurodevelopmental disorders. Int J Mol Sci. 2020;21:9303.
58. Inza I, Calvo B, Armananzas R, Bengoetxea E, Larranaga P, Lozano JA. Machine learning: an indispensable tool in bioinformatics. Bioinforma Methods Clin Res. Totowa, NJ: Humana Press; 2009. p. 25–48.
59. Lee HY, Jung SE, Lee EH, Yang WI, Shin KJ. DNA methylation profiling for a confirmatory test for blood, saliva, semen, vaginal fluid and menstrual blood. Forensic Sci Int Genet. 2016;24:75–82.
60. Hu W, Guan L, Li M. Prediction of DNA methylation based on multi-dimensional feature encoding and double convolutional

fully connected convolutional neural network. PLoS Comput Biol. 2023;19:e1011370.
61. Levy JJ, Titus AJ, Petersen CL, Chen Y, Salas LA, Christensen BC. MethylNet: an automated and modular deep learning approach for DNA methylation analysis. BMC Bioinformatics. 2020;21:1–15.
62. Macartney-Coxson D, Cameron AM, Clapham J, Benton MC. DNA methylation in blood—potential to provide new insights into cell biology. PLoS ONE. 2020;15:e0241367.
63. Campagna MP, Xavier A, Lechner-Scott J, Maltby V, Scott RJ, Butzkueven H, et al. Epigenome-wide association studies: current knowledge, strategies and recommendations. Clin Epigenetics. 2021;13:1–24.
64. Vanclay F, Baines JT, Taylor CN. Principles for ethical research involving humans: ethical professional practice in impact assessment Part I. Impact Assess Proj Apprais. 2013;31:243–53.
65. Wilhelm-Benartzi CS, Koestler DC, Karagas MR, Flanagan JM, Christensen BC, Kelsey KT, et al. Review of processing and analysis methods for DNA methylation array data. Br J Cancer. 2013;109:1394–402.
66. Weinhold L, Wahl S, Pechlivanis S, Hoffmann P, Schmid M. A statistical model for the analysis of beta values in DNA methylation studies. BMC Bioinformatics. 2016;17:1–11.
67. Di Lena P, Sala C, Prodi A, Nardini C. Missing value estimation methods for DNA methylation data. Bioinformatics. 2019;35:3786–93.
68. Price EM, Robinson WP. Adjusting for batch effects in DNA methylation microarray data, a lesson learned. Front Genet. 2018;9:338551.
69. Forest M, O'Donnell KJ, Voisin G, Gaudreau H, MacIsaac JL, McEwen LM, et al. Agreement in DNA methylation levels from the Illumina 450k array across batches, tissues, and time. Epigenetics. 2018;13:19–32.
70. Sun Z, Chai HS, Wu Y, White WM, Donkena KV, Klein CJ, et al. Batch effect correction for genome-wide methylation data with Illumina Infinium platform. BMC Med Genomics. 2011;4:1–12.
71. Zhuang J, Widschwendter M, Teschendorff AE. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. BMC Bioinformatics. 2012;13:1–14.
72. Affinito O, Palumbo D, Fierro A, Cuomo M, De Riso G, Monticelli A, et al. Nucleotide distance influences co-methylation between nearby CpG sites. Genomics. 2020;112:144–50.
73. Lövkvist C, Dodd IB, Sneppen K, Haerter JO. DNA methylation in human epigenomes depends on local topology of CpG sites. Nucleic Acids Res. 2016;44:5123–32.
74. Blattler A, Farnham PJ. Cross-talk between site-specific transcription factors and DNA methylation states. J Biol Chem. 2013;288:34287–94.
75. Maros ME, Capper D, Jones DT, Hovestadt V, von Deimling A, Pfister SM, et al. Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data. Nat Protoc. 2020;15:479–512.
76. Doherty T, Dempster E, Hannon E, Mill J, Poulton R, Corcoran D, et al. A comparison of feature selection methodologies and learning algorithms in the development of a DNA methylation-based telomere length estimator. BMC Bioinformatics. 2023;24:178.
77. Tang X, Mo Z, Chang C, Qian X. Group-shrinkage feature selection with a spatial network for mining DNA methylation data. Comput Biol Med. 2023;154:106573.
78. Arbet J, Yamaguchi TN, Shiah Y-J, Hugh-White R, Wiggins A, Oh J, et al. The Landscape of Prostate Tumour Methylation. bioRxiv. 2025;2025–02.
79. Passemiers A, Tuveri S, Sudhakaran D, Jatsenko T, Laga T, Punie K, et al. MetDecode: methylation-based deconvolution of cell-free DNA for noninvasive multi-cancer typing. Bioinformatics. 2024;40:btae522.
80. Mallik S, Seth S, Bhadra T, Zhao Z. A linear regression and deep learning approach for detecting reliable genetic alterations in cancer using dna methylation and gene expression data. Genes. 2020;11:931.
81. Sun H, Wang S. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. Bioinformatics. 2012;28:1368–75.
82. Lee D, Koo B, Kim S, Byun J, Hong J, Shin D-Y, et al. Increased local DNA methylation disorder in AMLs with DNMT3A-destabilizing variants and its clinical implication. Nat Commun. 2025;16:560.
83. Roessler J, Ammerpohl O, Gutwein J, Hasemeier B, Anwar SL, Kreipe H, et al. Quantitative cross-validation and content analysis of the 450k DNA methylation array from Illumina. Inc BMC Res Notes. 2012;5:1–7.
84. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics Proteomics. 2018;15:41–51.
85. Lucain M, Vitobello A, Sadikovic B, Albuisson J, Gaudillat L, Chevarin M, et al. Abnormal DNA methylation profile suggests the extension of the clinical spectrum of the SETD2-related disorders to a syndromic multiple tumor phenotype. Am J Med Genet A. 2025;197:e64043.
86. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46:389–422.
87. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
88. Quraishi BM, Zhang H, Everson TM, Ray M, Lockett GA, Holloway JW, et al. Identifying CpG sites associated with eczema via random forest screening of epigenome-scale DNA methylation. Clin Epigenetics. 2015;7:1–11.
89. Duckett D, Vormittag-Nocito ER, Jamshidi P, Sukhanova M, Parker S, Brat DJ, et al. Accurate identification of primary site in tumors of unknown origin (TUO) using DNA methylation. NPJ Precis Oncol. 2025;9:8.
90. Sill M, Schrimpf D, Patel A, Sturm D, Jäger N, Sievers P, et al. Advancing CNS tumor diagnostics with expanded DNA methylation-based classification. medRxiv. 2025;2025–05.
91. Papanicolau-Sengos A, Aldape K. DNA methylation profiling: an emerging paradigm for cancer diagnosis. Annu Rev Pathol Mech Dis. 2022;17:295–321.
92. Musolf AM, Holzinger ER, Malley JD, Bailey-Wilson JE. What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics. Hum Genet. 2022;1–14.
93. Ran R, Wang M, Miao J. The prognosis prediction model for endometrial cancer based on DNA methylation signature. Cancer Rep. 2025;8:e70218.
94. Yuan T, Edelmann D, Moreno V, Georgii E, de E Sousa ALB, Pelin H, et al. Identification and external validation of tumor DNA methylation panel for the recurrence risk stratification of stage II colon cancer. Transl Oncol. 2025;57:102405.
95. Yuan T, Edelmann D, Fan Z, Alwers E, Kather JN, Brenner H, et al. Machine learning in the identification of prognostic DNA methylation biomarkers among patients with cancer: a systematic review of epigenome-wide studies. Artif Intell Med. 2023;143:102589.
96. Kalyakulina A, Yusipov I, Bacalini MG, Franceschi C, Vedunova M, Ivanchenko M. Disease classification for whole-blood DNA methylation: meta-analysis, missing values imputation, and XAI. Gigascience. 2022;11:giac097.
97. Kwon Y, Blazyte A, Jeon Y, Kim YJ, An K, Jeon S, et al. Identification of 17 novel epigenetic biomarkers associated with anxiety disorders using differential methylation analysis followed by machine learning-based validation. Clin Epigenetics. 2025;17:24.
98. Liu Y, Taha HB, Zhang Q, Pan Z, Chatzipantsiou C, Wade E, et al. NANOME: A Nextflow pipeline for haplotype-aware allele-specific consensus DNA methylation detection by nanopore long-read sequencing. bioRxiv. 2025;2025–06.
99. Zarean E, Li S, Wong EM, Makalic E, Milne RL, Giles GG, et al. Evaluation of agreement between common clustering strategies for DNA methylation-based subtyping of breast tumours. Epigenomics. 2025;17:105–14.
100. Schwendinger S, Jaschke W, Walder T, Hench J, Vogi V, Frank S, et al. DNA methylation array analysis identifies biological subgroups of cutaneous melanoma and reveals extensive differences with benign melanocytic nevi. Diagnostics. 2025;15:531.
101. Fernández L, Pérez M, Orduña JM, Alcaraz JM. A new dimensionality reduction technique based on the wavelet transform for cancer classification. J Big Data. 2025;12:9.
102. Bahado-Singh RO, Ibrahim A, Al-Wahab Z, Aydas B, Radhakrishna U, Yilmaz A, et al. Precision gynecologic oncology: circulating cell free DNA epigenomic analysis, artificial intelligence and the accurate detection of ovarian cancer. Sci Rep. 2022;12:18625.
103. Colacino A, Soricelli A, Ceccarelli M, Affinito O, Franzese M. Subtypes detection of papillary thyroid cancer from methylation assay via deep neural network. Comput Struct Biotechnol J. 2025. https://doi.org/10.1016/j.csbj.2025.04.034.
104. Shao D, Addagudi S, Cowles J, Jain A, D'Souza L, Gore S, et al. Rarenet: a deep learning model for rare cancer diagnosis. Sci Rep. 2025;15:22732.

105. Yu X, Long H, Zeng R, Zhang G. Multi-task encoder using peripheral blood DNA methylation data for Alzheimer's disease prediction. Electronics. 2025;14:2655.
106. Bedi P, Rani S, Gupta B, Bhasin V, Gole P. Epibrcan-lite: a lightweight deep learning model for breast cancer subtype classification using epigenomic data. Comput Methods Programs Biomed. 2025;260:108553.
107. Jeong Y, Gerhäuser C, Sauter G, Schlomm T, Rohr K, Lutsik P. Methylbert enables read-level DNA methylation pattern identification and tumour deconvolution using a transformer-based model. Nat Commun. 2025;16:788.
108. Yassi M, Chatterjee A, Parry M. Application of deep learning in cancer epigenetics through DNA methylation analysis. Brief Bioinform. 2023;24:bbad411.
109. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. 2017. p. 618–26.
110. Yuan D, Jugas R, Pokorna P, Sterba J, Slaby O, Schmid S, et al. crossNN is an explainable framework for cross-platform DNA methylation-based classification of tumors. Nat Cancer. 2025;1–12.
111. Zhao X, Sui Y, Ruan X, Wang X, He K, Dong W, et al. A deep learning model for early risk prediction of heart failure with preserved ejection fraction by DNA methylation profiles combined with clinical features. Clin … [Internet]. 2022; Available from: https://link.springer.com/article/https://doi.org/10.1186/s13148-022-01232-8
112. Sokolov AV, Schiöth HB. Decoding depression: a comprehensive multi-cohort exploration of blood DNA methylation using machine learning and deep learning approaches. Transl Psychiatry. 2024;14:287.
113. Raiaan MAK, Sakib S, Fahad NM, Al Mamun A, Rahman MA, Shatabda S, et al. A systematic review of hyperparameter optimization techniques in convolutional neural networks. Decis Anal J. 2024;11:100470.
114. Yousefi PD, Suderman M, Langdon R, Whitehurst O, Davey Smith G, Relton CL. DNA methylation-based predictors of health: applications and statistical considerations. Nat Rev Genet. 2022;23(6):369–83.
115. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, ... TensorFlow: Large-scale machine learning on heterogeneous systems. 2015;
116. Paszke A, Gross S, Massa F, Lerer A, ... Pytorch: An imperative style, high-performance deep learning library. Adv Neural … [Internet]. 2019; Available from: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html
117. Pedregosa F, Varoquaux G, Gramfort A, Michel V, ... Scikit-learn: Machine learning in python journal of machine learning research. J Mach Learn …. 2011;
118. Kuhn M. Building predictive models in R using the caret package. J Stat Softw [Internet]. 2008; Available from: https://www.jstatsoft.org/index.php/jss/article/view/v028i05
119. Lang M, Binder M, Richter J, Schratz P, ... mlr3: A modern object-oriented machine learning framework in R. J Open … [Internet]. 2019; Available from: https://doi.org/10.21105/joss.01903.pdf
120. Dean J, Corrado G, Monga R, Chen K, Devin M, Mao M, et al. Large scale distributed deep networks. Adv Neural Inf Process Syst. 2012;25.
121. Jouppi N, Young C, Patil N, Patterson D. In-datacenter performance analysis of a tensor processing unit. Proc 44th … [Internet]. 2017; Available from: https://doi.org/10.1145/3079856.3080246
122. Buitinck L, Louppe G, Blondel M, Pedregosa F. API design for machine learning software: experiences from the scikit-learn project. ArXiv Prepr ArXiv … [Internet]. 2013; Available from: https://arxiv.org/abs/1309.0238
123. Huber W, Carey V, Gentleman R, Anders S. Orchestrating high-throughput genomic analysis with Bioconductor. Nat … [Internet]. 2015; Available from: https://www.nature.com/articles/nmeth.3252/1000
124. Zheng SC, Breeze CE, Beck S, Dong D, Zhu T, Ma L, et al. Epidish web server: epigenetic dissection of intra-sample-heterogeneity with online GUI. Bioinform. 2020. https://doi.org/10.1093/bioinformatics/btz833.
125. Kaushal A, Zhang H, Karmaus WJ, Ray M, Torres MA, Smith AK, et al. Comparison of different cell type correction methods for genome-scale epigenetics studies. BMC Bioinform. 2017;18:216.
126. Catoni M, Tsang JM, Greco AP, Zabet NR. DMRcaller: a versatile R/bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. Nucleic Acids Res. 2018;46:e114–e114.
127. Hackenberg M, Barturen G, Carpena P, Luque-Escamilla PL, Previti C, Oliver JL. Prediction of CpG-island function: CpG clustering vs. sliding-window methods. BMC Genomics. 2010;11:327.
128. Huang H, Chen Z, Huang X. Age-adjusted nonparametric detection of differential DNA methylation with case-control designs. BMC Bioinformatics. 2013;14:86.
129. Zaghlool SB, Al-Shafai M, Al Muftah WA, Kumar P, Falchi M, Suhre K. Association of DNA methylation with age, gender, and smoking in an Arab population. Clin Epigenetics. 2015;7:6.
130. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021;14:49–58.
131. Villicaña S, Bell JT. Genetic impacts on DNA methylation: research findings and future perspectives. Genome Biol. 2021;22:127.
132. Cheng Y, Gadd DA, Gieger C, Monterrubio-Gómez K, Zhang Y, Berta I, et al. Development and validation of DNA methylation scores in two European cohorts augment 10-year risk prediction of type 2 diabetes. Nat Aging. 2023;3:450–8.
133. Fan S, Tang J, Li N, Zhao Y, Ai R, Zhang K, et al. Integrative analysis with expanded DNA methylation data reveals common key regulators and pathways in cancers. NPJ Genom Med. 2019;4:2.
134. Esteller M. Epigenetics in cancer. N Engl J Med. 2008;358:1148–59.
135. Modhukur V, Sharma S, Mondal M, Lawarde A, Kask K, Sharma R, et al. Machine learning approaches to classify primary and metastatic cancers using tissue of origin-based DNA methylation profiles. Cancers. 2021;13:3768.
136. Zhou X, Cheng Z, Dong M, Liu Q, Yang W, Liu M, et al. Tumor fractions deciphered from circulating cell-free DNA methylation for cancer early diagnosis. Nat Commun. 2022;13:7694.
137. Zhang Z, Lu Y, Vosoughi S, Levy JJ, Christensen BC, Salas LA. HiTAIC: hierarchical tumor artificial intelligence classifier traces tissue of origin and tumor type in primary and metastasized tumors using DNA methylation. NAR Cancer. 2023;5:zcad017.
138. Yang Y, Zeng Q, Liu G, Zheng S, Luo T, Guo Y, et al. Hierarchical classification-based pan-cancer methylation analysis to classify primary cancer. BMC Bioinformatics. 2023;24:465.
139. Orozco JI, Knijnenburg TA, Manughian-Peter AO, Salomon MP, Barkhoudarian G, Jalas JR, et al. Epigenetic profiling for the molecular classification of metastatic brain tumors. Nat Commun. 2018;9:4627.
140. Liu MC, Oxnard GR, Klein EA, Swanton CSMCC, Seiden MV, Liu MC, et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol. 2020;31:745–59.
141. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. Neuro-oncol. 2021;23:1231–51.
142. Jurmeister P, Bockmayr M, Seegerer P, Bockmayr T, Treue D, Montavon G, et al. Machine learning analysis of DNA methylation profiles distinguishes primary lung squamous cell carcinomas from head and neck metastases. Sci Transl Med. 2019;11:eaaw8513.
143. Koelsche C, Schrimpf D, Stichel D, Sill M, Sahm F, Reuss DE, et al. Sarcoma classification by DNA methylation profiling. Nat Commun. 2021;12:498.
144. Jurmeister P, Glöß S, Roller R, Leitheiser M, Schmid S, Mochmann LH, et al. DNA methylation-based classification of sinonasal tumors. Nat Commun. 2022;13:7148.
145. Kuschel LP, Hench J, Frank S, Hench IB, Girard E, Blanluet M, et al. Robust methylation-based classification of brain tumours using nanopore sequencing. Neuropathol Appl Neurobiol. 2023;49:e12856.
146. Vermeulen C, Pagès-Gallego M, Kester L, Kranendonk M, Wesseling P, Verburg N, et al. Ultra-fast deep-learned CNS tumour classification during surgery. Nature. 2023;622:842–9.
147. Brändl B, Steiger M, Kubelt C, Rohrandt C, Zhu Z, Evers M, et al. Rapid brain tumor classification from sparse epigenomic data. Nat Med. 2025;1–9.
148. Patel A, Göbel K, Ille S, Hinz F, Schoebe N, Bogumil H, et al. Prospective, multicenter validation of a platform for rapid molecular profiling of central nervous system tumors. Nat Med. 2025;1–11.
149. Nabais MF, Laws SM, Lin T, Vallerga CL, Armstrong NJ, Blair IP, et al. Meta-analysis of genome-wide DNA methylation identifies shared associations across neurodegenerative disorders. Genome Biol. 2021;22:1–30.

150. Park C, Ha J, Park S. Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. Expert Syst Appl. 2020;140:112873.
151. Chen L, Saykin AJ, Yao B, Zhao F, Alzheimer's Disease Neuroimaging Initiative (ADNI). Multi-task deep autoencoder to predict Alzheimer's disease progression using temporal DNA methylation data in peripheral blood. Comput Struct Biotechnol J. 2022;20:5761–74.
152. Karaglani M, Agorastos A, Panagopoulou M, Parlapani E, Athanasis P, Bitsios P, et al. A novel blood-based epigenetic biosignature in first-episode schizophrenia patients through automated machine learning. Transl Psychiatry. 2024;14:257.
153. Gan Y, Yue W, Sun J, Yang D, Fang C, Zhou Z, et al. Exploration of epigenetic mechanisms and biomarkers among patients with very-late-onset schizophrenia-like psychosis. Neuropsychiatr Dis Treat. 2025. https://doi.org/10.2147/NDT.S513992.
154. Tian R, Zhang H, Wang C, Zhou S, Zhang L, Wang H. Research on prediction of multiple degenerative diseases and biomarker screening based on DNA methylation. Int J Mol Sci. 2025;26:313.
155. Gunasekara CJ, Hannon E, MacKay H, Coarfa C, McQuillin A, Clair DS, et al. A machine learning case–control classifier for schizophrenia based on DNA methylation in blood. Transl Psychiatry. 2021;11:412.
156. Barbu MC, Shen X, Walker RM, Howard DM, Evans KL, Whalley HC, et al. Epigenetic prediction of major depressive disorder. Mol Psychiatry. 2021;26:5112–23.
157. Zhang W, Lukacsovich D, Young JI, Gomez L, Schmidt MA, Martin ER, et al. DNA methylation signature of a lifestyle-based resilience index for cognitive health. Alzheimers Res Ther. 2025;17:88.
158. Bahado-Singh RO, Radhakrishna U, Gordevičius J, Aydas B, Yilmaz A, Jafar F, et al. Artificial intelligence and circulating cell-free DNA methylation profiling: mechanism and detection of Alzheimer's disease. Cells. 2022;11:1744.
159. Bahado-Singh RO, Vishweswaraiah S, Turkoglu O, Graham SF, Radhakrishna U. Alzheimer's precision neurology: epigenetics of cytochrome P450 genes in circulating cell-free DNA for disease prediction and mechanism. Int J Mol Sci. 2023;24:2876.
160. Fu H, Huang K, Zhu W, Zhang L, Bandaru R, Wang L, et al. Circulating cell-free DNA methylation profiles as noninvasive multiple sclerosis biomarkers: A proof-of-concept study. medRxiv. 2025.
161. Wang N, Li S, Yang L. DNA methylation patterns and predictive models for metabolic disease risk in offspring of gestational diabetes mellitus. Diabetol Metab Syndr. 2025;17:147.
162. Nong Y, Huang H, Xu L, Tan X, Xu S, Zhou X, et al. Association of DNA methylation epigenetic markers with all-cause mortality and cardiovascular disease-related mortality in diabetic population: a machine learning-based retrospective cohort study. Diabetol Metab Syndr. 2025;17:221.
163. Zhang X, Wang C, He D, Cheng Y, Yu L, Qi D, et al. Identification of DNA methylation-regulated genes as potential biomarkers for coronary heart disease via machine learning in the Framingham Heart Study. Clin Epigenetics. 2022;14:122.
164. Dogan MV, Grumbach IM, Michaelson JJ, Philibert RA. Integrated genetic and epigenetic prediction of coronary heart disease in the Framingham Heart Study. PLoS ONE. 2018;13:e0190549.
165. Dunn CM, Sturdy C, Velasco C, Schlupp L, Prinz E, Izda V, et al. Peripheral blood DNA methylation–based machine learning models for prediction of knee osteoarthritis progression: biologic specimens and data from the Osteoarthritis Initiative and Johnston County Osteoarthritis Project. Arthritis Rheumatol. 2023;75:28–40.
166. Thompson M, Hill BL, Rakocz N, Chiang JN, Geschwind D, Sankararaman S, et al. Methylation risk scores are associated with a collection of phenotypes within electronic health record systems. NPJ Genom Med. 2022;7:50.
167. Desiderio A, Pastorino M, Campitelli M, Longo M, Miele C, Napoli R, et al. DNA methylation in cardiovascular disease and heart failure: novel prediction models? Clin Epigenetics. 2024;16:115.
168. Haghshenas S, Karimi K, Stevenson RE, Levy MA, Relator R, Kerkhof J, et al. Identification of a DNA methylation episignature for recurrent constellations of embryonic malformations. Am J Hum Genet. 2024. https://doi.org/10.1016/j.ajhg.2024.07.005.
169. Turinsky AL, Choufani S, Lu K, Liu D, Mashouri P, Min D, et al. EpigenCentral: Portal for DNA methylation data analysis and classification in rare diseases. Hum Mutat. 2020;41:1722–33.
170. Hildonen M, Ciolfi A, Ferilli M, Cappelletti C, Al Alam C, Amor DJ, et al. Biallelic loss-of-function variants in ZNF142 are associated with a robust DNA methylation signature affecting a limited number of genomic loci. Eur J Hum Genet. 2025;1–8.
171. Sadikovic B, Aref-Eshghi E, Levy M, Rodenhiser D. DNA methylation signatures in mendelian developmental disorders as a diagnostic bridge between genotype and phenotype. Epigenomics. 2019;5:563–75.
172. Kaplanis J, Samocha KE, Wiel L, Zhang Z, Arvai KJ, Eberhardt RY. Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature. 2020;586:757–62.
173. McConkey H, Sadikovic B. Epigenomic mechanisms and episignature biomarkers in rare diseases. Epigenetics Hum Dis. Academic Press; 2024. p. 1031–76.
174. Aref-Eshghi E, Rodenhiser DI, Schenkel LC, Lin H, Skinner C, Ainsworth P, et al. Genomic DNA methylation signatures enable concurrent diagnosis and clinical genetic variant classification in neurodevelopmental syndromes. Am J Hum Genet. 2018;102:156–74.
175. Aref-Eshghi E, Bend EG, Hood RL, Schenkel LC, Carere DA, Chakrabarti R, et al. BAfopathies' DNA methylation epi-signatures demonstrate diagnostic utility and functional continuum of Coffin-Siris and Nicolaides-Baraitser syndromes. Nat Commun. 2018;9:4885.
176. Aref-Eshghi E, Kerkhof J, Pedro VP, Barat-Houari M, Ruiz-Pallares N, Andrau JC, et al. Evaluation of DNA methylation episignatures for diagnosis and phenotype correlations in 42 Mendelian neurodevelopmental disorders. Am J Hum Genet. 2020;106:356–70.
177. Hop PJ, Zwamborn RA, Hannon EJ, Dekker AM, van Eijk KR, Walker EM. Cross-reactive probes on Illumina DNA methylation arrays: a large study on ALS shows that a cautionary approach is warranted in interpreting epigenome-wide association studies. NAR Genomics Bioinforma. 2020;2:lqaa105.
178. Husson T, Lecoquierre F, Nicolas G, Richard AC, Afenjar A, Audebert-Bellanger S, et al. Episignatures in practice: independent evaluation of published episignatures for the molecular diagnostics of ten neurodevelopmental disorders. Eur J Hum Genet. 2023;1–10.
179. Levy MA, McConkey H, Kerkhof J, Barat-Houari M, Bargiacchi S, Biamino E. Novel diagnostic DNA methylation episignatures expand and refine the epigenetic landscapes of Mendelian disorders. Hum Genet Genom Adv. 2022. https://doi.org/10.1016/j.xhgg.2021.100075.
180. Schenkel LC, Schwartz C, Skinner C, Rodenhiser DI, Ainsworth PJ, Pare G, et al. Clinical validation of fragile X syndrome screening by DNA methylation array. J Mol Diagn. 2016;18:834–41.
181. Geysens M, Huremagic B, Souche E, Breckpot J, Devriendt K, Peeters H, et al. Clinical evaluation of long-read sequencing-based episignature detection in developmental disorders. Genome Med. 2025;17:1.
182. Ferraro F, Drost M, van der Linde H, Bardina L, Smits D, de Graaf BM, et al. Training with synthetic data provides accurate and openly-available DNA methylation classifiers for developmental disorders and congenital anomalies via MethaDory. medRxiv. 2025;2025–03.
183. Imperiale TF, Ransohoff DF, Itzkowitz SH, Levin TR, Lavin P, Lidgard GP, et al. Multitarget stool DNA testing for colorectal-cancer screening. N Engl J Med. 2014;370:1287–97.
184. Church T, Wandell M, Lofton-Day C, Mongin S, ... Prospective evaluation of methylated SEPT9 in plasma for detection of asymptomatic colorectal cancer. Gut [Internet]. 2014; Available from: https://gut.bmj.com/content/63/2/317.short
185. Taryma-Leśniak O, Sokolowska KE, Wojdacz TK. Current status of development of methylation biomarkers for in vitro diagnostic IVD applications. Clin Epigen. 2020;12:1–16.
186. Stewart G, Neste LV, Delvenne P, ... Clinical utility of an epigenetic assay to detect occult prostate cancer in histopathologically negative biopsies: results of the MATLOC study. J … [Internet]. 2013; Available from: https://www.auajournals.org/doi/abs/https://doi.org/10.1016/j.juro.2012.08.219
187. Jr RW, Neste LV, Moses K, ... Evaluation of an epigenetic assay for predicting repeat prostate biopsy outcome in African American men. Urology [Internet]. 2019; Available from: https://www.sciencedirect.com/science/article/pii/S0090429518303157

188. Shao S, Allen B, Clement J, Chung G, Gao J, ... Multi-cancer early detection test sensitivity for cancers with and without current population-level screening options. Tumori. 2023. https://doi.org/10.1177/03008916221133136
189. Bredno J, Lipson J, Venn O, Aravanis A, Jamshidi A. Clinical correlates of circulating cell-free DNA tumor fraction. PLoS ONE. 2021. https://doi.org/10.1371/journal.pone.0256436.
190. Wade R, Nevitt S, Liu Y, Harden M, Khouja C, Raine G, et al. Multi-cancer early detection tests for general population screening: a systematic literature review. Health Technol Assess Winch Engl. 2025;29:1.
191. Larson M, Pan W, Kim H, Mauntz R. A comprehensive characterization of the cell-free transcriptome reveals tissue-and subtype-specific biomarkers for cancer detection. Nat … [Internet]. 2021; Available from: https://www.nature.com/articles/s41467-021-22444-1
192. Dickson J, Bhamani A, Quaife S, Horst C, Tisi S. The reporting of pulmonary nodule results by letter in a lung cancer screening setting. Lung Cancer [Internet]. 2022; Available from: https://www.sciencedirect.com/science/article/pii/S0169500222004111
193. Chen X, Gole J, Gore A, He Q, Lu M, Min J, et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nat … [Internet]. 2020; Available from: https://www.nature.com/articles/s41467-020-17316-z.
194. Wu H, Guo S, Liu X, Li Y, Su Z, He Q, et al. Noninvasive detection of pancreatic ductal adenocarcinoma using the methylation signature of circulating tumour DNA. BMC Med. 2022. https://doi.org/10.1186/s12916-022-02647-z.
195. Gao Q, Lin Y, Li B, Wang G, Dong L, Shen B, et al. Unintrusive multi-cancer detection by circulating cell-free DNA methylation sequencing (THUNDER): development and independent validation studies. Ann … [Internet]. 2023; Available from: https://www.sciencedirect.com/science/article/pii/S092375342300087X
196. Kennedy E, Durm G, Farlow J. Multicancer early detection tests: a state-of-the-art review for otolaryngologists. OTO Open. 2024. https://doi.org/10.1002/oto2.70040.
197. Moran S, Martínez-Cardús A, Sayols S. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. Lancet … [Internet]. 2016; Available from: https://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(16)30297-2/abstract
198. Davalos V, Esteller M. Cancer epigenetics in clinical practice. CA Cancer J Clin. 2023;73:376–424.
199. Nishitani S, Smith AK, Tomoda A, Fujisawa TX. Data science using the human epigenome for predicting multifactorial diseases and symptoms. Epigenomics. 2024. https://doi.org/10.2217/epi-2023-0321.
200. Smits DJ, Debuy C, Brooks AS, Schot R, Ferraro F, Rots D, et al. Clinical utility of DNA-methylation signatures in routine diagnostics for neurodevelopmental disorders. Eur J Hum Genet. 2025;1–9.
201. Douillard J, Oliner K, Siena S. Panitumumab–FOLFOX4 treatment and RAS mutations in colorectal cancer. Engl J. 2013. https://doi.org/10.1056/NEJMoa1305275.
202. Mok T, Wu Y, Ahn M, Garassino M. Osimertinib or platinum-pemetrexed in EGFR T790M–positive lung cancer. Engl J. 2017. https://doi.org/10.1056/NEJMoa1612674.
203. Swain S, Baselga J, Kim S, Ro J. Pertuzumab, trastuzumab, and docetaxel in HER2-positive metastatic breast cancer. N Engl J. 2015. https://doi.org/10.1056/NEJMoa1413513.
204. Cassidy S, Schwartz S, Miller J, Driscoll D. Prader-willi syndrome. Genet Med [Internet]. 2012; Available from: https://www.sciencedirect.com/science/article/pii/S1098360021032846
205. Margolis S, Sell G, Zbinden M, Bird L. Angelman syndrome. Neurotherapeutics [Internet]. 2015; Available from: https://www.sciencedirect.com/science/article/pii/S1878747923016094
206. Tsurusaki Y, Okamoto N, Ohashi H, Kosho T, Imai Y. Mutations affecting components of the SWI/SNF complex cause Coffin-Siris syndrome. Nat … [Internet]. 2012; Available from: https://www.nature.com/articles/ng.2219
207. Witjes J, Morote J, Cornel E, Gakis G. Performance of the bladder EpiCheck™ methylation test for patients under surveillance for non–muscle-invasive bladder cancer: results of a multicenter …. Eur Urol … [Internet]. 2018; Available from: https://www.sciencedirect.com/science/article/pii/S2588931118300968
208. D'Andrea D, Soria F, Zehetmayer S, Gust K. Diagnostic accuracy, clinical utility and influence on decision-making of a methylation urine biomarker test in the surveillance of non-muscle-invasive bladder cancer. BJU … [Internet]. 2019; Available from: https://bjui-journals.onlinelibrary.wiley.com/doi/abs/https://doi.org/10.1111/bju.14673
209. Laukhtina E, Shim S, Mori K, David D, Soria F. Diagnostic accuracy of novel urinary biomarker tests in non–muscle-invasive bladder cancer: a systematic review and network meta-analysis. Eur Urol … [Internet]. 2021; Available from: https://www.sciencedirect.com/science/article/pii/S2588931121001838
210. Chen K, Kang G, Zhang Z, Lizaso A, Beck S, Lyskjaer I, et al. Individualized dynamic methylation-based analysis of cell-free DNA in postoperative monitoring of lung cancer. BMC Med. 2023. https://doi.org/10.1186/s12916-023-02954-z.
211. Vonk C, Hinai AA, Hanekamp D, Valk P. Molecular minimal residual disease detection in acute myeloid leukemia. Cancers [Internet]. 2021; Available from: https://www.mdpi.com/2072-6694/13/21/5431
212. Loomans-Kropp H. Multicancer early detection tests: where are we? JNCI Cancer Spectr [Internet]. 2023; Available from: https://academic.oup.com/jncics/article-abstract/7/1/pkac084/6858475
213. Dias S, Liu Y, Palmer S, Soares MO. A Bayesian approach to sharing information on sensitivity of a Multi-Cancer Early Detection test across and within tumour types and stages. ArXiv Prepr ArXiv250421517. 2025;
214. Neal R, Johnson P, Clarke C, Hamilton S, Zhang N, Cell-free DNA–based multi-cancer early detection test in an asymptomatic screening population (NHS-Galleri): design of a pragmatic, prospective randomised …. Cancers [Internet]. 2022; Available from: https://www.mdpi.com/2072-6694/14/19/4818
215. Gao Z, Liu Q, Zeng W, Jiang R, Wong WH. EpiGePT: a Pretrained Transformer model for epigenomics. bioRxiv. 2024;2023.07.15.549134.
216. Martínez-Enguita D, Dwivedi SK, Jörnsten R, Gustafsson M. NCAE: data-driven representations using a deep network-coherent DNA methylation autoencoder identify robust disease and risk factor signatures. Brief Bioinform. 2023;24:bbad293.
217. Steyaert S, Verhelle A, Criekinge WV. Variational autoencoders to predict DNA-methylation age and provide biological insights in age-related health and disease. medRxiv. 2023;2023.07.07.23292381.
218. Zeng W, Gautam A, Huson DH. Mulan-methyl—multiple transformer-based language models for accurate DNA methylation prediction. Gigascience. 2023;12:giad054.
219. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022;23:40–55.
220. Wekesa J, Kimwele M. A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment. Front Genet. 2023. https://doi.org/10.3389/fgene.2023.1199087/full.
221. Taha M, El-Hadidi M, Fawzi S. Deep learning applications in single-cell multi-omics analysis: a review. 2024 6th Nov Intell … [Internet]. 2024; Available from: https://ieeexplore.ieee.org/abstract/document/10753202/
222. Ballard J, Wang Z, Li W, Shen L, Long Q. Deep learning-based approaches for multi-omics data integration and analysis. BioData Min. 2024. https://doi.org/10.1186/s13040-024-00391-z.
223. Mohammed M, Abdulkareem K, Dinar A, Rise of deep learning clinical applications and challenges in omics data: a systematic review. Diagnostics [Internet]. 2023; Available from: https://www.mdpi.com/2075-4418/13/4/664
224. Clark C, Palta P, Joyce C, Scott C, Grundberg E. A comparison of the whole genome approach of MeDIP-Seq to the targeted approach of the infinium humanmethylation450 beadchip® for methylome. PLoS ONE. 2012. https://doi.org/10.1371/journal.pone.0050233.
225. Guanzon D, Ross J, Ma C, Berry O, Liew Y. Comparing methylation levels assayed in GC-rich regions with current and emerging methods. BMC Genom. 2024. https://doi.org/10.1186/s12864-024-10605-7.
226. Peters T, Meyer B, Ryan L, Achinger-Kawecka J. Characterisation and reproducibility of the HumanMethylationEPIC v2.0 BeadChip for DNA methylation profiling. BMC Genom. 2024. https://doi.org/10.1186/s12864-024-10027-5.
227. Elmarakeby HA, Hwang J, Arafeh R, Crowdis J, Gang S, Liu D, et al. Biologically informed deep neural network for prostate cancer discovery. Nature. 2021;598:348–52.

228. Prosz A, Pipek O, Börcsök J, Palla G, Szallasi Z, Spisak S, et al. Biologically informed deep learning for explainable epigenetic clocks. Sci Rep. 2024;14:1306.
229. Yang T-H, Yu Y-H, Wu S-H, Zhang F-Y. CFA: An explainable deep learning model for annotating the transcriptional roles of cis-regulatory modules based on epigenetic codes. Comput Biol Med. 2023;152:106375.
230. Consens ME, Dufault C, Wainberg M, Forster D, Karimzadeh M, Goodarzi H, et al. To Transformers and Beyond: Large Language Models for the Genome. ArXiv [Internet]. 2023;abs/2311.07621. Available from: https://api.semanticscholar.org/CorpusID:265158147
231. Hajjar M, Albaradei S, Aldabbagh G. Machine learning approaches in multi-cancer early detection. Information. 2024;15:627.
232. Pollard VT, Ryan MW, Mohanty A. FDA Issues Good Machine Learning Practice Guiding Principles. J Robot Artif Intell Law. 2022;5.
233. Ferber D, El Nahhas OS, Wölflein G, Wiest IC, Clusmann J, Leßman M-E, et al. Autonomous artificial intelligence agents for clinical decision making in oncology. ArXiv Prepr ArXiv240404667. 2024;
234. Canty RB, Bennett JA, Brown KA, Buonassisi T, Kalinin SV, Kitchin JR, et al. Science acceleration and accessibility with self-driving labs. Nat Commun. 2025;16:3856.
235. Ahani Azari A, Amanollahi R, Jafari Jozani R, Trott DJ, Hemmatzadeh F. High-resolution melting curve analysis: a novel method for identification of *Mycoplasma* species isolated from clinical cases of bovine and porcine respiratory disease. Trop Anim Health Prod. 2020;52:1043–7.
236. Wojdacz TK, Dobrovic A. Methylation-sensitive high resolution melting (MS-HRM): a new approach for sensitive and high-throughput assessment of methylation. Nucleic Acids Res. 2007. https://doi.org/10.1093/nar/gkm013.
237. Licchesi JD, Herman JG. Methylation-specific PCR. DNA Methylation Methods Protoc. 2009;305–23.
238. Kurdyukov S, Bullock M. DNA methylation analysis: choosing the right method. Biology. 2016;5:3.
239. Delaney C, Garg SK, Yung R. Analysis of DNA methylation by pyrosequencing. Immunosenescence Methods Protoc. 2015;249–64.
240. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the infinium methylation 450k technology. Epigenomics. 2011;3:771–84.
241. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010;466:253–7.
242. Qiu H, Sarathy A, Schulten K, Leburton J-P. Detection and mapping of DNA methylation with 2D material nanopores. Npj 2D Mater Appl. 2017;1:3.
243. Booth MJ, Ost TW, Beraldi D, Bell NM, Branco MR, Reik W, et al. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. Nat Protoc. 2013;8:1841–51.
244. Li G, Liu Y, Zhang Y, Kubo N, Yu M, Fang R, et al. Joint profiling of DNA methylation and chromatin architecture in single cells. Nat Methods. 2019;16:991–3.
245. Yan PS, Potter D, Deatherage DE, Huang TH-M, Lin S. Differential methylation hybridization: profiling DNA methylation with a high-density CpG island microarray. DNA Methylation Methods Protoc. 2009;89–106.
246. Irizarry RA, Ladd-Acosta C, Carvalho B, Wu H, Brandenburg SA, Jeddeloh JA, et al. Comprehensive high-throughput arrays for relative methylation (CHARM). Genome Res. 2008;18:780–90.
247. Gonzalgo ML, Liang G. Methylation-sensitive single-nucleotide primer extension (Ms-SNuPE) for quantitative measurement of DNA methylation. Nat Protoc. 2007;2:1931–6.
248. da Silva DAV, Brendebach H, Grützke J, Dieckmann R, Soares RM, de Lima JTR, et al. MALDI-TOF MS and genomic analysis can make the difference in the clarification of canine brucellosis outbreaks. Sci Rep. 2020;10:19246.
249. Ehrich M, Nelson MR, Stanssens P, Zabeau M, Liloglou T, Xinarianos G, et al. Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. Proc Natl Acad Sci. 2005;102:15785–90.
250. Sulewska A, Niklinska W, Kozlowski M, Minarowski L, Naumnik W, Niklinski J, et al. Detection of DNA methylation in eucaryotic cells. Folia Histochem Cytobiol. 2007;45:315–24.
251. Brena RM, Auer H, Kornacker K, Hackanson B, Raval A, Byrd JC, et al. Accurate quantification of DNA methylation using combined bisulfite restriction analysis coupled with the Agilent 2100 Bioanalyzer platform. Nucleic Acids Res. 2006;34:e17–e17.
252. Ca E. Combined bisulfite restriction analysis (COBRA). DNA methylation protocols. Methods Mol Biol. 2002;200:71–85.


## Publisher's Note