**Prediction of DNA N4 Methyl-Cytosine in six Organisms using Deep Learning**

# **Chapter 2:**

# **Related work and Literature Review:**

In this second chapter, we review the related work and the existing literature of our research topic DNA N4 methyl cytosine (4mC) site prediction. Section 2.1 describes the relation between computer sciences in the field of Bioinformatics. The significance of Artificial intelligence in bioinformatics is mentioned in section 2.2. Section 2.3 presents the application of computer science in bioinformatics in current era. To understand the epigenetic modifications, a brief discussion is given in section 2.4. Further, section 2.5 gives a brief overview of the development process in DNA N4 methyl cytosine prediction. Then the maturity of Deep learning in epigenetics is introduced in section 2.6. A detailed systematic literature review is presented in the next section 2.7. Research gaps and the Research model development are outlined in section 2.8 and 2.9 respectively. Lastly, chapter summary is stated in section 2.10.

## **2.1. Computer Science and Bioinformatics:**

“A scientific discipline that incorporates all characteristics of biological evidence, processing, storage, distribution, investigation, and explanation that combines the tools and practices of mathematics, computer science, and biology with the objective of reasoning the biological implication of a variety of data” is how bioinformatics is defined [1]. Any product that stocks, categorizes estimates, incorporates, examines, and/or allocates biological data is said to be a bioinformatics tool. After the origination of computerized protein and DNA sequencing technology in the mid-1970s, scholars began to utilize computers as central sequence archives that can be retrieved remotely; an advancement that occurred in the mid-1980s, bioinformatics gained popularity [2].

The practice of computer technology for the supervision and regulation of biological data is known as Bioinformatics. The formation of software tools will help in contributing in creation of appropriate knowledge within the biological area is one of the principal endeavors in the field of bioinformatics. Bioinformatics is distinct from the area of biological computing, despite the fact that they may interact. Bioinformatics employs computers for better understanding of biological problems, whereas computational biology is a field of computer engineering and computer sciences that influences biology and bioengineering for the evolution of biological computers.

Bioinformatics is an undeveloped and fast growing topic that bounced from the areas of investigational molecular biology and biochemistry, as well as the computer science disciplines of artificial intelligence, databases, and algorithms. To increase research in the domain of bioinformatics, one option is to tutor computer science to biologists and biotechnologists, while the other possibility is to train biology to computer scientists. The scientists have faith in in both methodologies, and are occupied on these issues to the best of their abilities. It is relatively easier to teach biologist’s novel computer technology. One of the main purposes of the bioinformatics industry is to magnify understanding about various biological processes. Bioinformatics contrasts from other approaches in achieving this objective because it emphasizes on the progress and application of specific procedures that are stated as "computationally intensive" by some [3].

Imagining, machine learning procedures, data mining, and pattern identification are examples of these approaches. Gene discovery, alignment of sequences, drug discovery and its design, protein structure prediction, connections among the proteins, and its alignment, gene expression prediction, and the modelling of development are just a few of the key scientific projects that have been done in this discipline. Some other examples for the exploration of sophisticated biological systems include computation regarding the development, discrete evaluations, statistical and probabilistic methods, molecular advancement, population heredities, and even forensic DNA analysis [4].

Rapid progressions in genomic exploration technology have arisen during the past several years. Additionally, enhancements in computer science technology have led to the creation of enormous amount of data in the domain of biology. Bioinformatics now includes the formation and development of procedures, databases, statistics, and mathematical approaches. This field also comprises the application of theory to hands-on and formal difficulties rising from the research and supervision of biological data.

Biologists cope with three types of data structures [5]:

* **Strings**. In order to denote DNA, RNA and amino acid sequences.
* **Trees**. Helps to depict the development of different creatures; 3D point sets and their connections. To signify structures of protein.
* **Graphs**. To signify metamorphic and motioning trails.

Sub-strings, sub-trees, sub-sets of points and links, and subgraphs are also prevalent among biologists. Strings that comprises of words and expressions are also engaged to embody interpretations that interconnect a sense provided by academics, although these understandings are often uncertain and flawed. Biological data is commonly regarded as by its huge bulk, the frequency of experimental inaccuracies (noise), replication, and even impreciseness. A biologist must study three concerns when deducing purpose from accessible data:

* Genetic factor, or sub-strings of DNA suited for producing proteins
* Protein structures denoted in 3D spatial
* The roles of these proteins contained by metamorphic and motioning trails.

## **2.2. The Role of Bioinformatics:**

Processing biological data involves in generating the necessary data, which may also require in constructing and administrating software applications. These agendas may also utilize to:

* Artificial intelligence
* Data mining
* Graph theory
* Soft computing
* Simulation of computers
* Processing Images

With discrete mathematics, these algorithms may evolve to much more. Control theory, statistics, and information theory in computer science, bioinformatics is constructed using multiple procedures.

**Artificial Intelligence**

The word itself signifies that bioinformatics applications are consuming computers to mimic human intelligence. Artificial intelligence (AI) is a contemporary advance technology. In the same way, it describes to a variety of topics. For example, in bioinformatics for DNA sequencing and reconstruction. It also aids in the creation of tools required for data processing. Most researchers indicate molecular biology as "tailor-made" for AI and machine learning approaches. This is due to the nature of AI practices, which surpass in fields with a enormous amount of data but little theory. Several algorithms have been formulated and used on various data sets since the advent of AI in this bioinformatics. The majority of these studies compare an innovative style to established methodologies, checking their effectiveness and productivity in specific data sets. DNA sequencing is one of the most noteworthy molecular biology activities. DNA chips are assumed to provide a quicker substitute to the more typical gel-based sequencing technologies. [6].

**Data mining**

The technique of spontaneously determining distinctive and understandable models and patterns from massive volumes of data is known as data mining. Bioinformatics is the study of storing, investigating, and using evidence derived from biological data such as sequences, molecules, gene expressions, and trails. Accepting these unceasingly rising streams of biological data will need the expansion of fresh data mining tools [7].

Universally, data mining is crucial in a multiple fields, particularly prediction. For every single process, this demands classification and grouping procedures. For each one has its own fixed set of algorithms for recognizing and handling biological data in order to conduct biomedical literature searches. This distinctive approach benefits in the development of Meta probing for scholars by permitting them to access numerous web databases from a single location. These practices are obligatory for graph theory, data integration, and text mining.

In the domain of bioinformatics, data-mining technology based on ML is becoming increasingly significant. Bioinformatics has developed essential developments since it deals with gigantic amounts of biological data. The regulation mechanism of a typical illness may be examined from the perspective of the complete system by merging multi-level data from biological tests and resourcefully applying suitable data mining technologies. This is tremendously important in the field of life science [8].

**Graph Theory**

Graphs are useful in three different ways. Graphs offer a data structure for repressing knowledge. Graph representations of controlling, signal transduction, and metabolic systems are examples. This might be in the form of informal drawings like those that the circles and arrows found in computational biology textbooks, or more openly in databases like Reactome[9].

The modelling of measurable data is another use of graphs in molecular biology. Numerous kinds of molecular biology investigations produce data that discloses molecule interactions. The data in a Yeast-Two-Hybrid screen, for example, is an innovation that a couple of proteins cooperated to form a transcription commencement complex. The data in a chromatin immunoprecipitation microarray test (ChIP-chip) is the asset of the pulled down protein's binding to the inquired DNA regions, which may be associated to one or more genes whose transcription is measured by them. Graphs can also be used in statistical modelling. For example, given some data containing interpretations of pairwise interactions or protein co-precipitation, one could want to develop a standard that defines various collections of proteins may link together to create a protein compound [10].

The graph approach, in specific, is for evaluation. In this domain, graph theory is used to associate sequences. It also involves setting the portions together and covering the graphs for scientific processing. It provides a strong and easy-to-understand representation.

**Soft Computing**

An expert system is created by accumulating information from a group of professionals. A biologist can make a decision with the encouragement of an expert system. By picking rules, selecting benchmarks, and investigating a scenario, the expert system articulates a deduction. Soft computing methodologies can be used to overcome this experiment. Soft computing tools can mine these essentials and then vigor rules that are personalized to the expert's actions. One of the features of biological data is missing and noisy data. This is somewhat that outdated computer methodologies cannot accomplish. By picking hedges in the data, soft computing-based methodologies can deal with omitted and noisy data [11].

Soft computing has the capability to discover earlier undiscovered associations between gigantic amounts of continuously altering biological data. It is likely that the data comprises significant hidden associations and connections. Soft computing methodologies, which are manufactured to handle very vast data sets, can be used to mine these types of correlations. Fuzzy sets, Artificial Neural Networks, GAs, RSes, and Support Vector Machines are instances of soft computing examples that have been used to investigate protein sequences, arrangements and folds, microarrays, and regulatory networks. Soft computing offers for estimated, worthy answers rather than the high accuracy, globally ultimate solution, allowing for a quicker arrival to a low-cost goal. [12].

Furthermore, though many technologies are refining for storing biomedical data, computers continue to play a significant role and have a precise position among scholars and biologists. This is an exceptional method of articulating gene data. Moreover, it states bioinformatics material. This helps in progression with the use of artificial neural networks and neural network models. Similarly, for exploring the procedure, this is the easiest and most consistent way. Proteomic and genomic uses are the most important phases of this methodology. This is helpful for scientists who are doing studies that produce a large volume of data.

**Simulation of computers**

There are a numerous simulation methodologies available, but computer simulation is the most consistent, flexible, and portable technique for maintaining data. Bioinformatics applications, in particular, are proficient at producing large amounts of data. Simulation is a type of computing used for building algorithms and software, creating databases and curating, and evaluating sequences, functions, and other data.

To compute geographic variation in cancer risk, scholars makes use of global models with bivariate smoothing funcationality. Using a simulation implementation, Siangphoe and Wheeler estimated the effectiveness of countless smoothing functions in universal additive models to classify whole spatial discrepancy of threat and sensitive threat in multiple geographical locations[13].

Several bioinformatics research and development organizations are occupied on building simulation-based tools to identify if a potential chemical for a new cure would cause poisonousness in patients before money is spent on actually manufacturing the medicine. In this regard, simulation is a process of recognizing problem areas and safeguarding that all variables are stated prior to the start of building on the drug development facility. Simulations may be used to demonstrate why certain happenings occur, where inadequacies exist, and if certain system alterations would recompense for or eliminate these ineffectiveness [14].

**Processing Images**

Typically, this methodology is required at every level of the virtually development process. It benefits biologists and researchers in inspecting their bioinformatics study. This technique virtually portrays each level of implementation. In contrast to other approaches in computer science, it is a crucial one. It leads to the progression of visual communication trends.

Bioinformatics technologies make it possible to retrieve information that might assist a scientist during a diagnostic, allowing them to quickly discover problems and track their progression over time. Medical and biological pictures have rapidly increased in size and information richness in recent years. For example, the right interpretation of pictures may be critical for early illness identification. Image data handling is an important aspect in order to diagnose clinical and exchange of information regarding health [15].

All representations of biological material assimilated using medical ways such as CT or MRI are denoted as "bio images." Bioinformatics tactics enable the formation of explanations that help in capturing the images for analyzing and incorporating of biological datasets for example gene data and ontologies in order to detect patterns and illnesses. For example, intriguing topic of exploration in neuroimaging is the relationship between brain and cognitive procedures utilizing various approaches such as fMRI) The latter is based on the BOLD signal, that depicts an increase in blood oxygenation during activities (such as hand activities) and imitates local neural movement. This method inspects the motion of the hydrogen atoms that make up the water particles in the brain using the BOLD signal in fMRI [16].

**Benefits of Bioinformatics**

The strategies listed above are quite valuable for producing research findings, as well as for researchers. As a result, using computer science to bioinformatics offers a more resourceful way of assessing, disintegrating, and sequencing data.

## **2.3. Modern day application of Computer Science in Bioinformatics:**

Bioinformatics encompasses a variety of scientific curricula and incorporates a variety of aspects in genomic analysis, such as

1. Genomics computation (storing & exploration of DNA sequences by means of several software)
2. Proteomics (laboratory procedures that help to decide what kind of proteins are actually represented and their potential cellular location),
3. Functional genomics (combination of function prediction and experimental testing of biological matter)
4. Profiling transcription (investigation of mRNAs at several phases of cell development & circumstances)
5. Structure function determinations (both anticipated and determined 3-D structures and function assessments with identified protein structures and functions).

The genetic data that is extracted encodes several biological insights whose interpretation can be the root for remarkable scientific and marketable success, bioinformatics is encountering a high demand for rapid answers. Following the pre-genomic period, which was marked by the attempt to structure the human genome, comes the post-genomic period that emphases on acquiring the concealed information buried in the genetic text. Currently, every scientist looking into humans or any other intriguing being, whether microbe, plant, or animal, for whom genome sequence data is now accessible or will be accessible in the near future, is given all of the mandatory evidence on the entire DNA sequence. Bioinformatics research is projected to have a substantial influence on improving our understanding of a variety of subjects. The domain of bioinformatics is proposed to have a noteworthy impression on the international economy. The pharmacological industry and drug expansion, agricultural science, health care, and the environment have all aided from inquiries in this subdivision. Bioinformatics procedures and practices will increase our capability to cure genetic human illnesses and mature new human treatments. The advantages and application of bioinformatics to humanity are numerous. Various applications are stated below [17]:

**Health Care**

Bioinformatics is a new subdivision of biological investigation that is swiftly attaining power. Bioinformatics technologies are supporting in the administration and investigation of genetic and biological information for health care. Bioinformatics is a sub-discipline of Biological Informatics that deals with biomedical complications at the molecular level. Not only does it have a direct influence on the progress of new diagnosis, treatment, and vaccines, but it also has an impact on our interpretation of infectious disease appliances, pathogen-host communications, and transmittal phases. The accessibility of bioinformatics tools is assisting in comprehending the probable advantages of the human genetic development, such as the prediction of disease vulnerability genes and the progress of many new treatments, as well as the capability to forecast which patients are expected to have hostile reactions or who are likely to have better efficiency.

The use of a bioinformatics methodology to figure out proteomic data can lead to the proper association of clinical indicators of patient receptiveness to a particular treatment.

The detection of discrepancies in people's genetic structure will lead to improved familiarity of the genes that result or contribute to diseases. However, in the end it will contribute health care workers in recommending the suitable treatment at the appropriate quantity for each individual. Additionally, this knowledge might benefit in reducing the unwanted side effects of today's "one-size-fits-all" medication recommending technique. Bioinformatics field is expected to result in novel disease classifications. Illnesses that have traditionally been considered based on their indications (phenotype) may be categorized based on their genetic behaviors (genotype)

**Drug Detection**

Improvement in drug procedures is challenging, laborious, and expensive. A candidate medicine normally takes around 5 years to mature, with scientific stages progressing to likely commercial accessibility taking even longer, at a estimated budget of more than 700 million dollars. The pharmacological manufacturing's attention has transformed from experimental approach drug detection to a balanced, structure-based medication design. The time and cost of manufacturing sustainable pharmacological agents might be cut in half if an effective and consistent drug design approach is used.

The identification of likeliness of drug, is nothing more than the prediction and removal of candidate molecules that are doubtful to endure the following phases of detection and growth, is done using computational methods. With the help of algorithms of genetic nature, we might forecast drug-likeness. Bioinformatics is crucial for categorizing and endorsing innovative objectives in order to save money on laboratory resources. Genomics has turn itno a significant supply of pharmaceutical objectives, and bioinformatics is vital to classify and authenticate new objectives.

Bioinformatics may benefit these techniques by providing some purposeful evidence on target candidates and associating this understanding to organic pathways using frequently informed community databanks. Tools of bioinformatics can be used to unravel the working of drugs and their effectiveness, along with how disease pathophysiology works. Drug research has spawned a number of new technologies. BioSuite is wide-ranging bioinformatics software suite that mergers macromolecular sequence and structural analysis, chemo informatics, and algorithms that can assist in development of drugs.

DOVIS is a utility software for high-throughput simulated transmission that exploits AutoDock, is a built simulated docking molecule screening and is a fundamental technique in improvement of drugs that is used to radically diminish number of chemical compounds that needs examination.

**Forensic Enquiry**

Forensic DNA are mainly multidisciplinary fields that makes use of tools from statistics and computer science to biological and legal matters. Two chief topics of forensics DNA analysis are exclusive identity and relatedness to other people. Large numbers of biological models and DNA profiles must be organized, analyzed, and associated in mass calamity occurrences, requiring the usage of computational laboratory managing structures for full-scale model cataloguing and tracking, as well as bioinformatics tools for DNA database probing.

PCR intensification of STR and its use of bioinformatics tools such as GeneScan and GenoTyper for detecting the existence or deficiency of STR related with sample. Genomic sequences may be re-sequenced with the help of microarray and analyzed using bioinformatics customary methodologies whenever bacteria, insects, or plants need to be identified at a crime scene.

Forensic investigation of the World Trade Center tragedy, New York, USA, September 11, 2001, was conducted using bioinformatics techniques such as CODIS, DNA-View, MDKAP, and MFISys. Bioinformatics approaches that scan DNA databases were also engaged to help fatalities of the South Asian Tsunami tragedy.

**Crop Enhancement**

Over the years, people chose various plants according to their requirements & formed agricultural plants that have countless benefits in terms of value, capacity, and agricultural practices over natural (wild) plants. Nevertheless, due to so many qualities are involved in opposition and quality, refining them all at the same time is incredibly puzzling. The genomic uprising has made it imaginable to increase a widespread genetic knowledge of performance of plant at many stages of assembling.

Sophisticated organic procedures that build infection resistance structures and guarantee crop quality are now accessible for organized functional examination exploiting bioinformatics procedures. Genome agendas create an immense amount of data, that must be administered, stored, and dispersed to international scientific community. The objective of Plant bioinformatics include promising the public domain compliance of all sequencing data via databases, providing rational explanation of genetic, proteomic, and phenotypes, thus enlarging associations within and between plants' data as well as other organisms Currently, bioinformatics in crops is certainly essential in breeding plant. Merging modelling and virtual reality with candid breeding agendas, as well as normalizing data assembling for phenotypic, ecological, and genetic data, is paving the way for next-generation breeding of plants.

Bioinformatics usage includes analyzing sequences, transcriptomes, proteomes, and metabolomes. Through bioinformatics-based data assimilation, 1D–omics information such as genes, proteins, and metabolic that is transformed into cell reality. By the year 2050, bioinformatics will modernize the way biological research is performed. Every scientist will devote time to the computational research by using internet in order to produce and explain data and exploration of data, and determine the data of other people to associate. Furthermore, bioinformatics will enable to search for current information in the field, apply it to his or her findings, and issue his or her findings to rest of the world. Tools of Bioinformatics and datasets are created in order to construct an in-silico structure of the plant that includes stages of organization and can be further used to estimate system feedback.

**Investigation of Food**

Agronomic-ally basic crops such as maize, soybean, potato, and cotton have all been engineered using agricultural biotechnology methodologies. Executive consent is compulsory before such products may be retailed. Before commercialization, each product endures extensive food, feed, and environmental safety testing, which comprises sensitivity test for each new protein added in the genetically amended crop.

One of the central goals is avoid the initiation of allergens that are known or experimentally reactive partial allergy, so that food-allergic people are not open to upsetting foods or allergens when they are not expecting it. Algorithms of bioinformatics and dedicated databases collecting known as well as mistrusted allergens are critical and convenient for recognizing known and probably cross-reactive allergens that might signify a distress to people that have sensitivities.

If a considerable counterpart is found, the presented protein must be verified again using sera from people sensitive to the identical protein to perceive the probability that produce reactions. Databases of allergenic proteins, such as SDAP food, is used to aid in estimating the Immunoglobulin E-binding prospective of unique food proteins and relativities amongst known allergens. Nutrients have been proven to modify gene struture and the combination of abundant proteins at various points in the progressions that lead to enzymes, structural proteins, and other constituents that are obligatory for life. As an outcome, the quantity and even the kind of nutrients accessible through gene expression might have an impression on protein synthesis. As a result, fewer proteins are produced, or it is created in a less-than-ideal operational form, or no protein is formed at all.

Due to the hereditary nature of genes existing, as well as other genes that are normal or have polymorphisms that may influence the structure of gene, these possibilities may arise. Researchers are able to realize all biological bearings of a dietary factor, meal, or diet if they have enough statistics and the precise bioinformatics tools. By analyzing Single Nucleotide Polymorphisms, that specify particular mixture in genomes, bioinformatics approaches may be used to estimate a variety of problems independently, including alterations in defenselessness or virus risk, and healing effectiveness. Consequently, it is anticipated that in the upcoming years, more customized material on food functionality would be accessible for disorder prevention, in harmony with specific genome inconsistencies.

**Biodiversity Management**

Systematic knowledge on biodiversity can be found in a variation of databases, on paper, or through museums, and repositories. In most situations, material on organisms is not accessible, and therefore it is not exploited to make strategies or findings about them. Around the world, examiners are occupied to boost our capability to handle biodiversity documents, and the new domain of biodiversity informatics is pursuing to discourse the domain's concerns. Within the situation of biodiversity statistics, new procedures and tools magnifies existing computer engineering and informatics tactics.

Biological complication of animals, the occurrence of a vast quantity of species and environments, and the relations of species with many abiotic variables all subsidize to the intricacy of biodiversity information. For both the organization and long-term care of biological diversity evidence, good database structural design are necessary.

Biodiversity informatics demands structures and procedures that can handle extensive variety of biological data, from molecules to morphological behaviors, inhabitants, and surroundings.

Many countries have formed associations with the resolution of diversity of biology, and numerous have merged other contracts related to biodiversity maintenance. These contracts have unswerving significances in which diverse biological information is assembled, prepared, preserved and circulated within countries. Tools are needed to create networks between new datasets and outdated data, as well as data from disparate sources, such as ecosystem-scale universal change and carbon cycle data, collect data in novel ways, investigate as well as produce it, and show the outcome in an understandable and operative format.

For supervision and publicizing biodiversity data, technologies like as notebooks and DNA bar code systems have been produced. For unifying, sustaining, and observing biodiversity, software programs such as Plant Info, Garden Info, Seed Pack, Biolit, and databanks such as Fungal Database Meliolales, Sacred Groves of Kerala, and Wild Ornamental Plants of the Western Ghats have been created.

Disturbing species risk valuation tools may be recognized by increasing analytical biological models for all non-native classes in a given district and then employing these models to policy procedures to avert species assaults. The domain of biodiversity informatics, which is still in its early period, seems to be on the brink of making huge revolutions that might alter the Internet into a gigantic universal biodiversity information system.

**Other Applications:**

Genes and proteins have a probable to discover genes or loci that normalize economically vital features in animals. Amplified growth rate and improved cadaver structure, boosted feed consumption, changed milk arrangement, enriched mohair assembly, and superior generative presentation are all declared as transgenic alteration candidates.

The techniques of Bioinformatics have been engaged in veterinary exploration to discover emergent ailments and to form additional explanations for the remaining enhancement and improvement of molecular diagnostics. Bioinformatics is being used to create worldwide repositories in microbiology in order to create an incremental knowledge repository that allows a large amount of investigational data and meta-data about bacteria, as well as to progress wide-ranging data mining tools for awareness in this data-rich setting.

To create vigorously modernized and adjustable portals based on experimental bacterial variety and associated biotechnological improvements, with the vital objective of commercializing freshly exposed insights as new requests or products. Climate change research is also sponsored by bioinformatics. Many species depend on only on carbon dioxide that is cause of carbon, and intensifying carbon dioxide discharges are one of the key drivers of global climate transformation. The examination of these minute species' genetics, which is achievable thanks to bioinformatics, benefits in the improvement of policies to reduce carbon dioxide intensities.

## **2.4. Brief overview of Epigenetics Modifications:**

Few decades ago, noteworthy advancement has been achieved in the subject of epigenetics, particularly in the areas of DNA methylation, histone modification, and non-coding RNA. The investigation of how cells influences gene activity without transforming the DNA sequence is called epigenetics. In Greek, "epi-" denotes "on or above," therefore "epigenetic" refers to influences other than the genetic code. A methyl group interacts with a DNA nucleotide, resulting in DNA methylation, thanks to DNA methyl-transferases (DNMTs). Epigenetic modifications are changes to DNA that affect whether genes are turned on or off. These changes are produced to DNA, but they have no effect on the sequence of the DNA building blocks. Epigenetics is a collection of modifications that regulate the activity (expression) of genes across a cell's DNA (genome).

Epigenetic modifications have an effect on the formation of proteins in cells because they assist regulate whether genes are switched on or off. This mechanism makes sure that just the proteins necessary for the cell's function are formed. Proteins that excite bone formation, for example, are not produced in muscle cells. Individuals, tissues within an individual, and even cells within a tissue have diverse patterns of epigenetic alteration. Environmental aspects such as a person’s nourishment and exposure to impurities can confine the epigenetic. Epigenetic variations can be passed down from generation to generation and can even be passed down from parent to child.

The two forms of epigenetic modifications are [18]:

1. DNA methylation
2. Histone modification

**DNA methylation:**

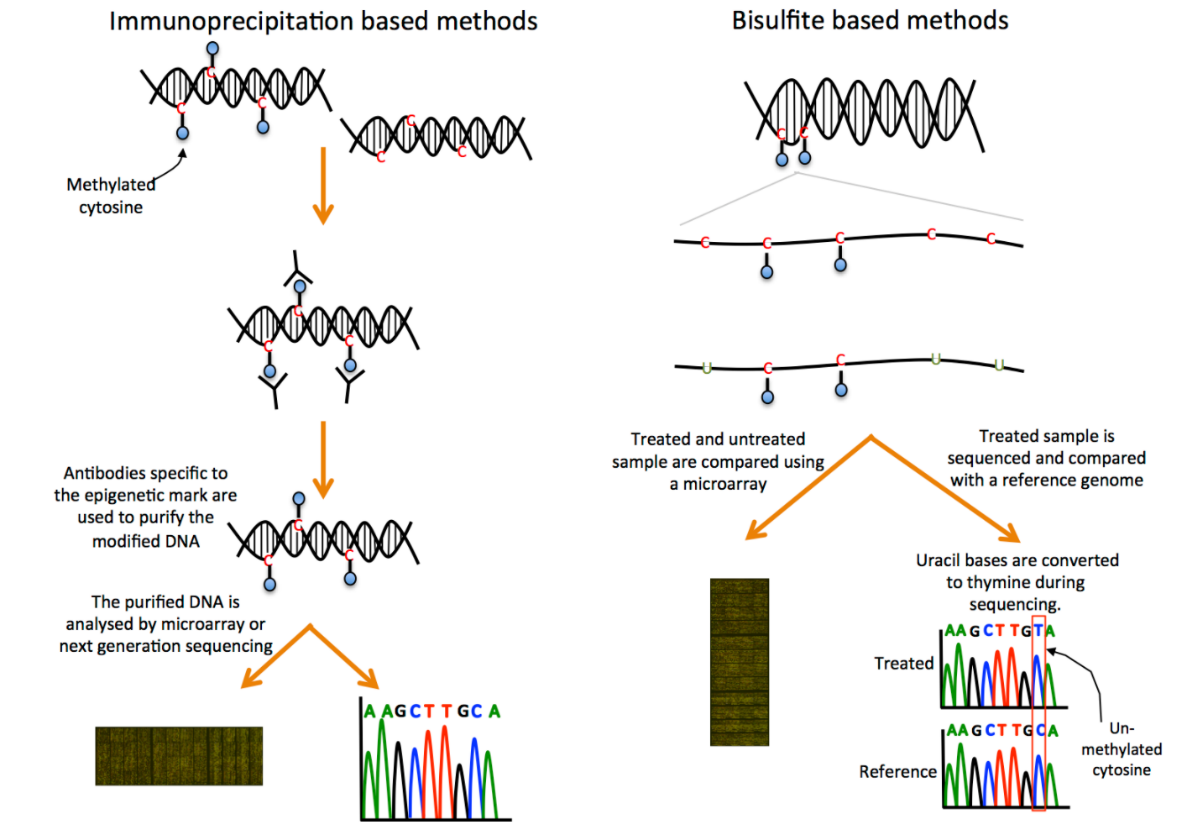
DNA methylation is an exceptional epigenetic modification seen in both prokaryotes and eukaryotes that is controlled by a specific methyl transferase enzyme. DNA methylation anomalies have been associated to cancer, asthma, and diabetes. DNA methylation is a widespread kind of epigenetic modification. The attachment of tiny chemical groups termed methyl groups, involving of one carbon atom and three hydrogen atoms, to DNA building blocks is known as DNA methylation. When methyl groups are attached on a gene, the gene is quietened or turned off, hence protein is not created. DNA methylation is an important epigenetic alteration that affects numerous biological processes such as, DNA conformation, genomic imprinting, cell growth, and the capability to alter DNA protein connections. It also supports transcription regulation, X chromosomal inactivation stabilization, and duplex stability.

There are three kinds of DNA methylation based on the state of methylation modification: N4-Metylcytosine (4mC), 5-Methylcytosine (5mC), and N6-Methyladenine (6mA). 5mC is the most predominant and well-studied epigenetic procedure in eukaryotic cells, in which methyl modification takes place on the fifth position of the cytosine pyrimidine ring. The majority of the study is focused on the 5mC sites, which are significant in a variation of biological procedures that can be activated by diabetes, cancer, and other neurological illnesses. When methylation occurs on the sixth position of the adenine purine ring, DNA 6mA is formed, which is found in many prokaryotic organisms. There is a large increase in genetic inconsistency because of these methyl alterations, and 6mA plays an important role in DNA duplication, gene rebuilding, and cell progression. DNA 5mC is also associated to a number of biological happenings, including transposon suppression, imprinting, and gene regulation.

Furthermore, methylation of the cytosine pyrimidine ring at the fourth position leads in DNA 4mC, which was later discovered in bacteria utilizing high-grade susceptible and high yielding methods. This is important in the Restriction-Modification (R-M) system because it protects host DNA from restriction enzyme degradation. DNA 4mC helps to monitor a variety of developments, including cell cycle, DNA replication mistakes, expression, and genetic cell stabilization. The 4mC sites are primarily responsible for checking, regulating, and correcting DNA replication mistakes. The methylation process appears to limit the Watson-Crick hydrogen bonding structure with guanosine, according to various species. The fact that self and foreign DNA have diverse tendencies means that in some operations, such as cytosine methylation, the nuclear DNA may be protected. Furthermore, the 4mC methylation protects its DNA by preventing enzyme-mediated destruction. Using chromatin immunoprecipitation (ChIP) or bisulfite-based techniques, DNA methylation may be determined.

**ChIP-based method:**

ChIP-based approaches filter methylated areas of the genome using methylation-specific antibodies. The DNA is then studied using microarrays or next-generation sequencing (NGS) to identify these areas.



**Bisulphite-based method:**

Bisulfite-based measures treat DNA samples with bisulfite, which alters un-methylated cytosine bases to uracil bases while keeping methylated residues as cytosine. The preserved DNA is then examined using microarrays or next-generation sequencing (NGS). Only methylation types sensitive to bisulfite-induced alterations may be perceived using this method.

The bisulfite-treated sample is crossbred with an unprocessed sample for microarray analysis. The signal intensity ratio amongst samples is employed to determine the quantity of methylation in particular genomic regions. NGS associates the sequences of the preserved and unprocessed samples to find precise methylation sites in the genome. A well-annotated genome is required for bisulfite sequencing.

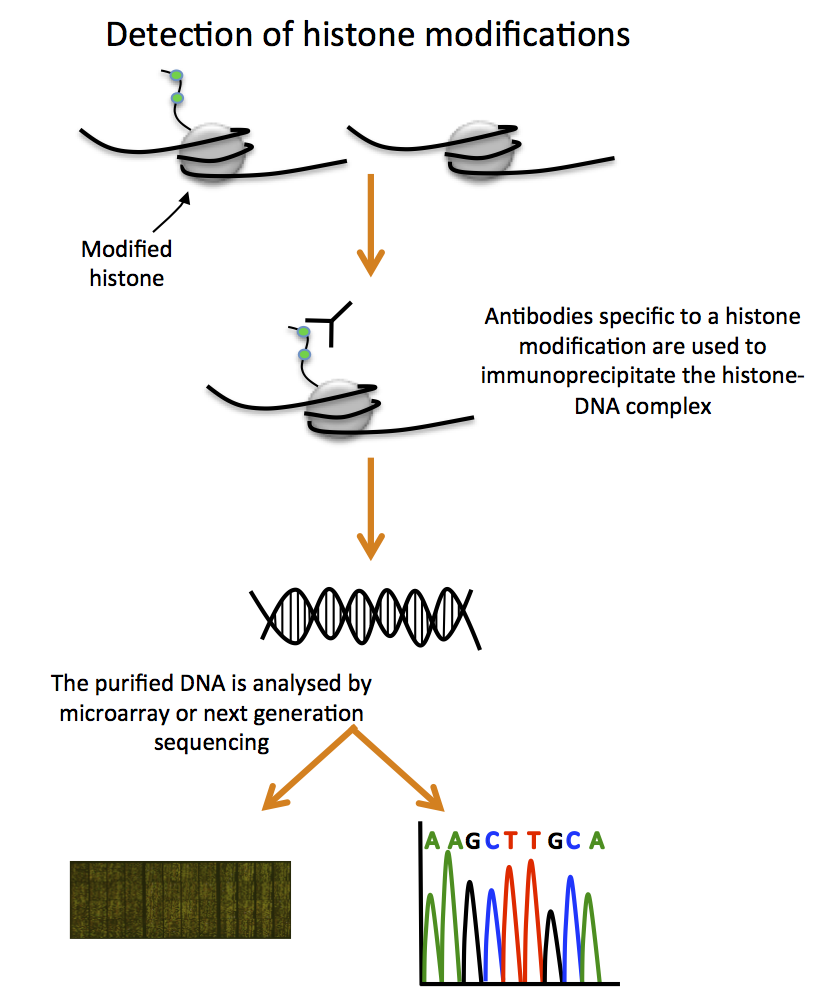
**Histone modification:**

Histones are structural proteins found in the nucleus of the cell. The structure of chromosomes is determined by the way DNA wraps around histones. Histones can be altered by adding or removing chemical groups such as methyl or acetyl groups (each comprising of two carbon, three hydrogen, and one-oxygen atoms). Acetylation, methylation, ubiquitination, citrullination, and phosphorylation of certain amino acids within the histone protein, generally towards the C-terminal ('tail') end of the protein, are all instances of histone modifications. By altering, the way histones attach to DNA, these changes can influence both positively and negatively on gene expression.

Histones are vulnerable to posttranslational modifications (PTMs), which are crucial constituents of governing circuits that control chromatin changing aspects and actions with the original DNA. Histone alterations are also epigenetic passengers that may be transferred along to daughter cells, permitting lineage-specific transcription forms to be maintained. As a result, accepting the roles of histone modifications has been a major emphasis of research in the chromatin area.

Histone changes can be recognized using a number of methods, including mass spectrometry and genomics techniques like ChIP-chip and ChIP-seq. To determine segments of the genome linked with these variations, genomics procedures combine chromatin immunoprecipitation (ChIP) of particular, modified histones and their supplementary DNA with microarray (chip) or next-generation sequencing (NGS) of the DNA molecules.

Errors in the epigenetic process can result in incorrect gene activity or inactivity, such as modifying the wrong gene or failing to add a chemical group to a specific gene or histone. Genetic abnormalities are frequently caused by altered gene activity, including that induced by epigenetic mistakes. Scientists are still studying the relationship between the genome and the chemical molecules that affect it. They are observing at the impressions of epigenetic changes and mistakes on gene function, protein engineering, and human health in particular.



## **2.5. Development Process of DNA N4 methyl cytosine:**

DNA methylation, in its different forms, has been linked to a wide range of biological activities in organisms from all over the world. N6-methyladenine (mA), for example, is originated largely in GATC sequence environments in particular bacteria and supports in the regulation of duplication, the disparity repair process, and the appearance of certain genes. In plants, 5-methylcytosine may be found in a variety of sequence settings, each of which is administered by various genetic practices. In vertebrates, mC is found typically in C-G dinucleotides, which usually clusters in CpG islands near or at transcription start sites. Methylation in CpG islands controls biological developments in cells and may also spread epigenetic inheritance to children [19].

Changes in mC arrangements are associated to cancer and other disorders and play a significant role in development. In human embryonic stem cells, ample cytosine methylation in non–C–G states was recently revealed, but not in segregated cells, indicating that it is a distinctive kind of methylation crucial in maintaining the pluripotent state [20], [21]. Finally, 5-hydroxymethylcytosine (hmC), which has been detected in neurons of mouse and embryonic stem cells, may be a freshly revealed epigenetic mark whose biological role is unidentified. Cytosine methylation is the most well-studied of the DNA modifications described above, owing to its importance in human health and disease. There's a lot of interest in mapping genome-wide mC forms across different cell types and in response to diverse environmental variables. [22].

Bisulfite treatment, which converts epigenetic data to genetic material by changing cytosine to uracil but not methyl cytosine, and massively parallel sequencing of DNA are presently the most widely used methods for cytosine methylation exploration. Researchers recently utilized this approach to create single-base resolution methylation maps across the human genome for the Arabidopsis thaliana genome, a portion of the mouse genome, and both fibroblasts and embryonic stem cells. Despite the advances made possible by bisulfite treatment–based sequencing, there are a number of faults in the method [23].

For example, sample preparation techniques connected with bisulfite treatment are typically costly and time-consuming, and the harsh reaction conditions necessary for complete conversion may damage DNA. [24] . Furthermore, altered genomes' decreased complexity restricts primer design for future PCR amplification and makes alignment to a reference genome more challenging. Finally, bisulfite-based sequencing can't tell the difference between cytosine, mC, and hmC. Thin layer chromatography, high-performance liquid chromatography, mass spectrometry, and Nano pore amperometry can all be used to detect methylation in nucleotides in solution [25], [26] .

To our knowledge, no high-throughput approach has been developed that allows primary sequence to be determined at the same time as methylation status. DNA methylation may be detected directly using single-molecule, real-time (SMRT) DNA sequencing. This method examines macromolecules' sequence and structure. Using this method, single DNA polymerase molecules may be seen in real time as they catalyze the integration of fluorescently tagged nucleotides matching to a template macromolecule strand. [25].

## **2.6. Deep Learning and its maturity in Bioinformatics:**

Deep learning (DL) has grown rapidly in bioinformatics, demonstrating a thrillingly promising ability to explore the intricate relationships masked in enormous biological and biomedical data. When compared to typical machine learning, deep learning is a relatively modern topic, and its use in bioinformatics is much novel. The recent comeback of artificial intelligence has been chiefly driven by machine learning. DL is the most important component of contemporary machine learning methodologies. Artificial neural networks (ANNs) are the stepping-stones of deep learning. They are demonstrated to be capable of approximating any nonlinear function within a definite accuracy and have been widely engaged to accomplish many computing problems [27].

Deep learning the subdivision of machine learning techniques that progressed from conventional NN’s through the elementary thought of neuron processors and the crucial structural design of multiple processing layers for the resolution of conveying relationships that are non-linear in nature, through reactions of each layer. It enables key and effective applications in a wide range of fields, including speech and image processing, handling of natural language, and biological investigation, thanks to its strong points in high-level concepts of enormous raw data attributes, distributed and parallel computing, and complex learning mechanisms that do not involve a lot of physical involvement. Deep learning (DL) methodologies developed as a subdivision of machine learning (ML) and are observed to be more competent and operational when allocating with enormous bulks of data, the age known as "big" data. These models have proven that they can reach better expectation accuracies than ever before [28].

Architectures of deep learning are (ANNs) artificial neural networks with numerous non-linear layers and different variants have been developed established on the attributes of the input data and the research goals. The fundamental disadvantage of ML over DL is that these algorithms cannot manipulate natural data in its raw form efficiently. DL has also demonstrated that it can provide models that are more precise and effectual in identifying patterns in high-dimensional data, making them beneficial to a wide range of applications. DL models, like ML models, require training data, but the quantity of data required in DL is more challenging and has a noteworthy impact on the expert model's prediction value. Minimums vary depending on the problem's complexity, but limited amount hundreds and thousands of instances is a decent starting point [29].

**Artificial Neural Networks:**

The neurons and networks that make up human brains provided the inspiration for Artificial Neural Networks (ANN). The ANN is made up of a set of fully linked nodes (neurons) that mimic the transmission of impulses from brain synapses -whether they fire or not- through the neural network. DL architectures like this are used for feature selection, classification, dimensionality reduction, or as a submodule of a more complex design like convolutional neural networks.

Similarly, an artificial neural network is made up of numerous computing elements known as neurons. These elements are interconnected together by links, each of which has a weight attached to it. Long-term memory is comparable to weights. Each element, like a real neuron, is input via input networks. Each element then calculates the biased total of the input values, and a transfer function interprets the ultimate result into the unit's output value[30].

ANNs are defect tolerance, which means they can handle noisy or fuzzy data as well as data that is partial or has absent values. They also have the skill to generalize, which means they can understand data that is not the identical as the training data. Thus, their capability to calculate future cases or tendencies founded on what they have formerly seen signifies a ‘real-world' explanation to a particular issue. As a result, trained ANNs can be used as separate operational structure in order to forecast the classification of an unidentified case of awareness, and thus have the probable to be used in diagnosis [31].

**Convolutional Neural Network (CNN):**

The Convolutional Neural Network (CNN) is a deep neural network architecture that is most typically used to analyze visual data. It was proposed as an entirely computerized image exploration system for classifying handwritten characters. CNNs are completely inter-connected networks based on the multilayer perceptron’s approach, in which each node/neuron in one layer is subsequently linked to all nodes in the next layer. A system of linked and tunable units that can transmit a signal from one unit to another is referred to as an ANN.

Convolutional Neural Networks have made advances in a variety of pattern recognition areas during the last decade, from image processing to speech recognition. The greatest advantage of CNNs is that they reduce the number of parameters in ANNs. This achievement has prompted academics and inventors to investigate larger models in order to solve perplexing problems that were previously difficult to solve with traditional ANNs; the most notable theory relating to CNN questions is that they should not include spatially dependent characteristics. In other words, in a face detection app, we do not need to be troubled with the position of the faces in the photos. The only thing that matters is that they be identified, irrespective of where they are in the representation. Another key attribute of CNN is the capability to mine features that are abstract in nature when input transmits deeper levels. In image recognition, for instance, the edges may be identified in the first layers, trailed by simpler shapes in the second layers, and finally sophisticated level features such as faces in the third layer [32].

The architecture of CNN comprises three layers that are[33][34]:

* Convolutional layers
* Pooling layers
* Fully connected (FC) layer

When these layers are stacked, a CNN architecture will be created. The dropout layer and the activation function are two more key aspects.

**Convolutional layer:**

This is the first layer, which extracts different characteristics from the input images. The convolution scientific procedure is performed between the input image and a NxN-size filter. By gliding the filter across the input picture, the dot product between the filter and the sectors of the input picture in terms of the filter's size is obtained (NxN). The result is the Feature map, which contains information about the image such as its corners and edges. This feature map is then passed on to subsequent layers, which use it to learn a variety of characteristics from the input image.

**Pooling layer:**

After a Convolutional Layer, a Pooling Layer is usually added. The main goal of this layer is to decrease computing expenses by reducing the extent of the convolved feature map. This is accomplished by reducing the network between layers and allowing each feature map to act independently. Depending on the method used, there are several types of pooling processes. The feature map in Max Pooling yields the largest element. Average Pooling is used to calculate the average of the elements in an encoded sized Image segment. The entire sum of the components in the marked section is estimated using Sum Pooling. The Convolutional Layer and the FC Layer are linked by the Pooling Layer.

**Fully Connected layer:**

The weights and biases, neurons, is what makes up the Fully Connected (FC) layer. Its purpose is to link the neurons between two layers. The last several layers of a CNN Architecture are generally situated before the output layer. The preceding layers' inputs the images and are flattened, supplied to the FC layer. The flattened vector is then sent to higher FC levels, where the functional operations are carried out. At this stage, the categorization process begins.

**Dropout layer:**

The training dataset is likely to overfit if all of the features are connected to the FC layer. When a model performs so well on training data that it has a negative impact on its performance when applied to new data, this is known as overfitting.

To solve this problem, a dropout layer is used, in which a few neurons from the neural network are disconnected during the training process, resulting in a smaller model. Following a dropout of 0.3, 30% of the nodes in the neural network are randomly dropped out.

**Activation Function:**

Finally, one of the most important components of the CNN model is the activation function. They're used to learn and estimate any type of network variable-to-variable relationship, no matter how long or intricate it is. To put it another way, it controls which model information should be sent forward and which should not at the network's end.

It creates non-linearity in the network. Some of the most commonly used activation functions include ReLU, Softmax, tanH, and sigmoid. Each of these functions has a certain purpose. The sigmoid and softmax functions are preferred for binary classification CNN models, although softmax is generally used for multi-class classification.

The basic idea behind this deep architecture is to calculate and absorb massive amounts of feature maps in order to infer non-linear connections between the input signal and the intended output. CNN is frequently used in image classification for feature extraction, selection, and reduction.[29].

**Recurrent Neural Network (RNN):**

The functioning of recurrent neural networks (RNN) is similar to that of normal feedforward neural networks (FNN) [13], in which nodes form a directed graph along a temporal sequence. RNNs may now demonstrate temporal dynamic behavior while also integrating internal memory. Recurrent networks can recall information from previously studied states thanks to their short-term memory, making them ideal for sequential signal processing and prediction models. The ability of RNNs to relate information from a prior job to the current job is one of its advantages.

RNN is a discriminative model that has also been engaged as a generative model, in which the model's "output" results denotes the expected input data. When an RNN is utilized as a discriminative model, the model's output result is given a label that matches to the input data [35].

**Long short-term memory (LSTM):**

Long short-term memory (LSTM) is a kind of RNN that can learn long-term dependencies and is specifically intended to avoid the problem of long-term dependencies. A LSTM unit has a cell/node, an input gate, an output gate, and a forget gate at its heart. The node considers values during particular times, while the input/output gates control the information flow.

Long short-term memory (LSTM) allows systems to retain inputs for a long time by employing an accumulator-like memory cell that has an association to itself at the next time reiteration and has a weight, and repeats its own real-valued state and sequential weights. Another unit that learns when to clear the content of the memory, multiplicatively gates this self-connection. LSTM networks have since shown to be more successful, particularly when many layers are used for each time step [35].

The basic concept of the LSTM cell is to organise its internal operations around two qualitatively different but complimentary goals: data and data control. The data modules create candidate data signals (from 1 to 1), while the control components prepare "throttle" signals (with values ranging from 0 to 1). The Long Short-Term Memory (LSTM) network's primary goal was to tackle the problem of vanishing gradients. A key breakthrough in the LSTM design was the incorporation of nonlinear, data-dependent controls into the RNN cell, which can be further trained and guarantees that the gradient of the objective function with respect to the state signal is retained.

## **2.7. Systematic Literature Review of DNA N4 methyl cytosine:**

### **2.7.1. Criteria for paper:**

In this section, we present the literature review. Following are the steps taken for gathering relevant literature for our research paper:

* **Searching Criteria:**

Our search criteria includes all papers from the year 2013 onwards. Online repositories such as IEEE, ACM, Google Scholar, Springer and Science Direct are part of literature review. Keywords like epigenetics, N4 methyl cytosine, methylation, deep learning, bioinformatics, etc. were taken into consideration for advance search to find relevant papers.

* **Inclusion Criteria:**

Inclusion Criteria of our paper includes those literature that is published in well-known journals that is ACM, IEEE, and Springer etc. are included as part of study. Thorough investigation of minimum 20 different research papers, are included based on search criteria that contains minimum 15 citations per paper.

* **Exclusion Criteria:**

Omission of Grey literature, such as presentations, reports, abstracts, unpublished papers, etc. from this study, while conducting literature review. Papers that are not in English language are excluded as well.

### **2.7.2 Related work:**

Single molecule real-time sequencing (SMRT) is a technique for detecting DNA methylation in order to investigate nucleic acid sequences and creation. Single DNA polymerase particles are tested in real time while catalysing the assimilation of fluorescently defined nucleotides matching to a template nucleic acid strand in this technique. Thousands of arrayed zero-mode waveguides (ZMWs), nanophotonic structures that reduce background fluorescence and enable for the use of large quantities of labelled nucleotides, which are necessary for rapid and efficient DNA sequencing by synthesis, are used to monitor these processes. When a nucleotide is integrated, it emits a pulse of fluorescence whose colour identifies the nucleotide. The fluorophore bound to the nucleotide's terminal phosphate is broken by the polymerase before it translocates to the next base in the DNA template. Typical SMRT sequencing synthesis rates in this technique are 1–3 bases per second.

A highly parallel confocal fluorescence detection device was used to acquire data. Previously, we described pulse calling, which utilised a threshold technique on dye-weighted fluorescence emission intensities, and read alignments, which utilised a Smith-Waterman method. Following alignment, reads were filtered to eliminate low-quality sequences obtained from doubly loaded ZMWs. IPD values were tallied from pairs of successfully aligned template locations and allocated to the pair's second template position. The duration of the pulses associated with correctly aligned base calls was used to calculate pulse width values. The smallest and greatest 5% of IPDs and pulse width values at each position were omitted from all analyses to eliminate outlier effects. When the DNA polymerase contacts m6A, m5C, or 5-hmC on the template strand during SMRT sequencing, significant variations in these kinetic parameters were found. The kind and location of the base change in the DNA template may be determined using these different kinetic fingerprints [37].

We use the SMRT sequencing approach to analyze MTase specificities by combining full DNA sequence determination with methylated base analysis in a single operation. We looked at 16 DNA substrates that were methylated in vivo by a variety of single prokaryotic MTases produced in an E. coli strain with no other MTase genes. The samples comprised MTases that introduced m6A, m4C, or m5C alterations, from MTases either with known substrate specificity or from those with unknown specificity. The findings revealed the MTase's extreme sequence specificity as well as the precise position of the methylated base [38].

Several DNAmethylation databases have been created, giving essential tools to the epigenetic community. MethDB[39] is the first database to record DNA methylation profiles and the gene expression data that goes with them. DNA methylation profiles acquired using the bisulfite sequencing approach are stored in NGSMethDB [40]. MethBank [41] is concerned with methylome alterations during embryonic development, whereas MethyCancer [42] and MENT [43] are concerned with malignancies. MethHC [44] is another database for the purpose of methylation and gene expression of human cancer. PubMeth [45] is a methylation database for cancer that is based on text mining of published literature. So far, none of these databases have contained DNA 5mC profiles, and none have supplied DNA 6mA or 4mC data.

The first database for DNA 6mA and 4mC methylomes, MethSMRT, was created using publically accessible SMRT sequencing data sets. The database contains 6mA and 4mC methylomes for 156 species, including seven eukaryotes and 149 prokaryotes, and may be used to host, analyse, browse, search, and download them. It also includes a genome browser for visualising methylation patterns and other data. Gene annotation and single nucleotide polymorphisms (SNPs). A brief summary is also provided by the database. Combining 6mA and 4mC methylome statistics and projected each species' methylation motifs. It is possible to compare methylomes between various species and data sets since all of the data sets in the database were processed using the same analytical pipeline with the same quality control. MethSMRT is freely available to the public at <http://sysbio.sysu.edu.cn/methsmrt/>. MethSMRT now has 156 species, including 7 eukaryotes and 149 prokaryotes, including Arabidopsis, C. elegans, Drosophila, Mouse, and Yeast [46].

To identify the 4mC sites, the authors suggested a support vector machine-based predictor namely iDNA4mC [47] in which DNA samples are created using nucleotide chemical property and nucleotide frequency. The MethSMRT database was used to get genomes. According to preliminary experiments, the best predictive results were obtained when the sequence length was 41 bp. As a result, the positive samples' sequences are all 41 bp. The suggested method's performance was evaluated using the jackknife test. In the jackknife test, each sample in the training dataset is selected as an independent test sample, and all attributes are computed without taking into account the one that is being identified. In iDNA4mC, three types of nucleotide chemical characteristics (ring structure, hydrogen bond, and chemical activity) were integrated in addition to sequence-based information by encoding DNA sequences using nucleotide frequency. The ring structure and hydrogen bond provide the highest contributions for 4mC site identification in both prokaryotes and eukaryotes, according to comparisons of the accuracies derived from the models based on different properties. iDNA4mC has 78.04 % accuracy in C. elegans, 81.16 % in D. melanogaster, 76.05 % in A. thaliana, 79.82 % in E. coli, 81.53 % in G. subterraneus, and 84.04 % in (G. pickeringii).

4mC predictor [48] incorporates Chou’s five step methodology [49] To encode sample sequences, they employed a system based on two approaches: position-specific trinucleotide sequence propensity (PSTNP) and EIIPs. Algorithm such as Naïve Bayes, KNN, Random Forrest, and SVM were incorporated in the study. Among all these machine learning algorithms SVM performed the best. According to jackknife testing, the 4mCPred I model trained with PSTNP features performed well when compared to an existing predictor, whereas the 4mCPred II model trained with hybrid features performed even better. The 4mC predictor outperformed IDNA4mC in terms of accuracies.

Authors [50] present a predictor namely 4mcPred-SVM and introduced a novel feature representation approach that combines the four sequence-based feature descriptors: (1) kmer dinucleotide frequency, (2) mono-nucleotide binary encoding, (3) dinucleotide binary encoding, and (4) local position-specific dinucleotide frequency in this paper. The same benchmark dataset of six organisms (C. elegans, D. melanogaster, A. thaliana, E. coli, G. subterraneous, and G. pickeringii) was employed in this study. Other than SVM, Naïve Bayes, KNN, Random Forrest, Gradient Boost and Logistic Regression were also implemented. The model SVM achieved the maximum accuracy of 0.8153, 0.8304, 0.7867, 0.8325, 0.8370, and 0.8603, respectively.

In this work, a new predictor called i4mC-ROSE [51] was proposed for detecting 4mC sites in the Rosaceae genomes of F. vesca and R. chinensis. Using random forest (RF) algorithms, six probabilistic sequence-encoding methodologies (KSNC, BE, EIIP, Kmer, DPCP, TPCP) were incorporated in this study. In order to improve prediction performance, the six statistical scores were integrated using a linear regression model. i4mC-ROSE is the first computational tool for predicting 4mC sites within two genomes Frag aria vesca and Rosa chinensis with an accuracy of 88% and 89% respectively.

The authors developed a 4mCPred-IFL predictor [52] employing an iterative feature representation technique for 4mC prediction utilising this novel methodology. BKF, PCP, MMI, EIIP, ring function hydrogen chemical characteristics, dinucleotide binary profile and frequency, Pseudo dinucleotide composition, and SVM are among the eight feature encoding techniques employed. Feature optimization, supervised probabilistic feature generation, and iterative feature generation are the three processes of the proposed iterative representation. Three methods are involved in the proposed 4mcPred-IFL prediction method. To begin, a 41-bp scan window is utilized to fragment genomic sequences into equal-length subsequences, with the ones centered on the C being kept and the others being eliminated. Second, the generated subsequences are encoded with features using iterative feature representation learning methodology. In the last stage, the characteristics for each sequence are put into the previously trained SVM model, which subsequently assigns a score to each sequence ranging from 0 to 1. If the score is more than 0.5, the sequence is a real 4mC site; otherwise, it is a non- 4mC site.

Furthermore, Meta-4mCpred [53], a predictor created by, combined six machine learning algorithms (SVM, RF, AB, GB, ERT, KNN) with seven feature encodings that encompass a wide range of sequence data, such as compositional, physicochemical, and NT position-specific data on the benchmark dataset of six organisms to provide a promising outcome. Then the authors used SVM to input these probabilistic characteristics created a final forecasting model Meta-analysis is used during cross-validation. With an average accuracy of 84.2 percent, 4mCpred came out on top.

Additionally, authors that employed an Ensemble Learning model [54] and incorporated the Chou’s five step rule for DNA N4 prediction in the mouse genome utilized four machine-learning algorithms (SVM, RF, GB, ERT).There are eight different types of feature encodings (BPF , EIIP, DPE, LPDF, kmer, RFHC, DPCP, TPCP) employed in this study. In addition, the authors developed 4mCpred-EL, a user-friendly online web server for quickly detecting 4mC sites in the mouse genome. These probabilistic characteristics were fed into four distinct machine learning (ML) classifiers, and matching models were created, which were then used to create an ensemble predictor that achieved an overall accuracy of 79.90 percent. This work, according to authors, is the first 4mC site prediction approach for the mouse genome. That will aid in the discovery of proper 4mC site predictions and might be applied to other species as well.

In this paper, the authors used Second-Order Markov Model (SOMM4mC) [55] to predict DNA N4 methyl cytosine sites in all six species of benchmark database MethSMART. SOMM4mC differs from prior 4mC prediction algorithms in that it takes into account the dependence between nearby nucleotides around the methylation site. Indeed, Markov chain models have long been used to leverage biological sequence adjacency dependence. SOMM4mC is the first tool for predicting 4mC sites based on adjacency dependency data. In six species, SOMM4mC and the first order Markov model beat the three current approaches (4mCPred-SVM, iDNA4mC, and 4mCPred), some with large accuracy gains. In comparison to first-order Markov findings, SOMM4mC has a superior classification performance. In dataset of E. Cole and C.elegans attained an accuracy of 91.8% and 87.6% respectively.

DNA4mC-LIP [56] is a new meta-predictor that improves the effectiveness of detecting 4mC sites by combining current models with a preliminary ideal weight. DNA4mC-LIP is, to our knowledge, the first classifier to include a combined technique in the prediction of 4mC sites. The DNA4mC-LIP was rigorously validated on separate datasets to demonstrate its performance. DNA4mC-LIP surpasses previous approaches for detecting 4mC sites, according on performance comparisons on separate datasets. iDNA4mC, 4mCPred, 4mcPred-SVM, 4mcPred-IFL, Meta-4mCpred, and 4mCpred-EL are all integrated into DNA4mC-LIP.

4mcDeep-CBI [57] is the first predictor that uses deep neural network in order to identify 4mC sites. The C.elegans dataset has been considerably enhanced, with the number of samples raised from 3108 to 17808, which will be useful for future study. 3-CNN and BLSTM are utilized to gather advanced features and extract deep information from the collected features. Experiments reveal that advanced features outperform traditional characteristics in detecting 4mC locations. Finally, we incorporate the probability feature matrix acquired by machine learning methods into the deep learning model, which improves prediction accuracy even more. In our trial, the proposed model's accuracy rose from 87 percent to 93 percent when compared to the state-of-the-art predictor.

Using convolution neural networks, a new sequence-based DNA 4mC sites predictor was constructed (4mCCNN) [58]. The grid search strategy was used to find the top performing hyper-parameters. In all species in the benchmark datasets, the suggested model outperformed the state-of-the-art approaches. A user-friendly web server was created and given accessible for free. Chou's five-step principles, which have been widely employed in the development of numerous predictors for proteome or genome analysis, were followed. The grid search technique was used to optimize the hyper parameters of 4mCCNN. The number of convolution layers, the number of filters in each convolution layer, the size of these filters, and the dropout rate are all parameters to consider. The suggested model 4mCCNN is a straightforward CNN with two one-dimensional convolution layers, 42 7-unit filters, and a single stride unit. A ReLU activation function and a dropout layer with a dropout rate of 0.6 follow each convolution layer. The CNN layers' learnt features are fed into a dropout layer with a rate of 0.5 and a fully connected layer with one node for prediction, followed by a sigmoid function.

DNC4mC-Deep [59], a new predictor, has been developed for the detection of 4mC sites in the freshly produced genomes of F. vesca, R. chinensis, and cross-species. The following is a summary of the overall structure of our research: first, we applied the six encoding techniques 2Kmer, 3Kmer, (BE), NCP, nucleotide chemical property, (NCPNF), and multivariate mutual information (MMI). Then, using the Convolution Neural Network, we created a deep learning model (CNN). To find the best model with tweaked hyper parameters, we used a grid search technique. The results of each encoding scheme were recorded and the K-fold cross-validation procedure with the value of K as 10 was used to input all six encoding methods independently in the optimally selected model. We used performance assessment measures to evaluate and analyze the model's findings on each encoding scheme. We also demonstrated two separate applications: the first is silico mutagenesis representation using heat maps, and the second is utilizing saliency maps to determine the most important regions of a sequence. After analyzing the outcomes of the model using all six distinct feature-encoding approaches, we discovered that DNC outperforms the state-of-the-art model in all six encoding schemes. DNC4mC-Deep obtains MCC of 0.635, 0.565, and 0.562 on F. vesca, R. chinensis, and cross-species independent datasets, respectively, in comparison to the state-of-the-art model.

Deep4mcPred [60] is a multilayer deep learning-based predictive model for detecting DNA N4-methylcytosine alterations, which we suggested. For the first time, we combine residual network and recurrent neural network, as well as an attention mechanism, to create a multi-layer deep learning prediction system in this predictor. Deep4mcPred incorporates a recurrent neural network for Long Short Term Memory (LSTM) and an attention mechanism into the Residual Networks (ResNet).The comparison findings show that our suggested model can detect 4mC locations more effectively than existing predictors can. The Deep4mcPred is employed on three benchmark species A.thaliana, C. elegans, and D.melanogaster. It contains 20,000 positive and negative samples. The neural network contained four layers and 89.3% accuracy was achieved in C.elegans, highest compared to other species. Furthermore, the suggested solution is implemented using a simple and user-friendly webserver that is publicly accessible. The major advantage this predictor has over other predictors is that while training, deep learning system does not require the features to be specified. It can learn high-level traits and capture the distinctive specificity of 4mC sites automatically, allowing it to differentiate authentic 4mC sites from non-4mC sites.

[61] Deep4mC, a new 4mC site predictor, was created utilizing deep convolutional neural networks (CNNs) and four representative characteristics. Deep4mC makes advantage of the feature assessment findings, which include binary, ENAC, EIIP, and NCP. For species with a tiny sample size. The authors added a bootstrapping mechanism to our deep learning framework to take use of the high number of negative samples (Ns) available to avoid false positives. Multiple CVs critically analysed Deep4mC's performance and compared it to existing approaches. In these species, the average AUC values of several CVs were all more than 0.9, with A. thaliana having the lowest (AUC= 0.9005) and E. coli having the highest (AUC= 0.9722). When compared to earlier methods in these six species, we discovered that Deep4mC could enhance AUC values from 10.14 to 46.21 percent using an independent dataset.

DeepTorrent [62] is a deep learning-based computer system for predicting 4mC sites from DNA sequence data. DeepTorrent, in particular, employs four distinct types of feature encoding algorithms to translate raw DNA sequences into input for deep networks, which include CNNs with inception (BLSTM) and an attention mechanism. To deal with the problem of limited sample sizes, it employs the deep transfer learning approach. When compared to state-of-the-art approaches, DeepTorrent gets the greatest performance for 4mC site prediction across all six investigated species in extensive benchmarking studies on two separate datasets.

|  |  |  |  |
| --- | --- | --- | --- |
| **Authors** | **Samples** | **Findings** | **Limitations** |
| [37] | - | Directly detect DNA methylation using Single Molecule, Real-Time DNA sequencing (SMRT) | 1. very costly 2. less efficient |
| [46] | 156 species | METHSMRT Database for 4mC and 6mA sites | 1. needs to be updated regularly |
| [47] | Benchmark dataset of six species (D. Melanogaster, C. Elegans, A. Thaliana, G. Subterraneous, E. coli, G. pickeringii) | Algorithm: SVM classifier Accuracy: 78.04%,76.05%,  81.16%, 84.04%,81.53%, 79.82% (respectively) | 1. Transcription factors and histone modifications are not incorporated in the study. |
| [48] | Benchmark dataset of six species (D. Melanogaster, C. Elegans, A. Thaliana, G. Subterraneous, E. coli, G. pickeringii) | Algorithm: Naïve Bayes, KNN, SVM, Random Forrest  Feature extraction technique: PSTNP and EIIP values | 1. Deep learning models with a hybrid feature approach not incorporated in the study. |
| [50] | Benchmark dataset of six species (D. Melanogaster, C. Elegans, A. Thaliana, G. Subterraneous, E. coli, G. pickeringii) | Algorithm: SVM  Four sequence-based feature descriptors and two-step optimization | 1. Combining sequence-based features with genomic features such as ChiP-seq, RNA-seq, and motif analysis approach not incorporated in the study. 2. Only one ML model included in the study. |
| [51] | Training dataset  F. vesca: 9708  R. Chinensis: 4674  Independent dataset  F. vesca: 3234  R. Chinensis: 1558 | Algorithm: Random Forrest  Six encoding techniques: KSNC, BE, EIIP, Kmer, DPCP, TPCP  Accuracy: 88% and 89% respectively | 1. ANOVA, mRMR, and MRMD not incorporated in the study. |
| [52] | Benchmark dataset of six species (D. Melanogaster, C. Elegans, A. Thaliana, G. Subterraneous, E. coli, G. pickeringii) | Algorithm: SVM  Eight various feature encoding techniques: BKF, PCP, MMI, EIIP, ring function hydrogen chemical properties, dinucleotide binary profile and frequency, Pseudo dinucleotide composition. | 1. Limited use of ML model i.e. only one ML model is implemented in the study. |
| [53] | Benchmark dataset of six species (D. Melanogaster, C. Elegans, A. Thaliana, G. Subterraneous, E. coli, G. pickeringii) | Algorithms: SVM, RF, AB, GB, ERT, KNN | 1. The methodology is generic can be enhanced to other sequence-based prediction problems 2. Limited use of genomic features |
| [54] | Mouse genome  Training set: 1600  Independent set: 360 | Algorithm: SVM, RF, GB, ERT, Ensemble Learning.  Accuracy: overall 79.90% | 1. Only sequence information opted in study 2. Informative features pseudo nucleotide composition and multivariate mutual information, deep features, and other ML models not included in the study. |
| [55] | Benchmark dataset of six species (D. Melanogaster,  C. Elegans, A. Thaliana, G. Subterraneous, E. coli, G. pickeringii) | Algorithm: Second-Order Markov Model  Accuracy: E. Cole and C.elegans 91.8% and 87.6% respectively | 1. The study can be enhanced by further integrating better features |
| [56] | Benchmark dataset of six species (D. Melanogaster, C. Elegans, A. Thaliana, G. Subterraneous, E. coli, G. pickeringii) | Integrated model of existing predictors  iDNA4mC, 4mCPred-Ⅰ, 4mCPred-Ⅱ, 4mCPred-SVM, 4mCPred-IFL, Meta-4mCPred | 1. The study uses limited predictors to develop an integrated model. |
| [57] | Benchmark dataset of six species (D. Melanogaster,  C. Elegans, A. Thaliana, G. Subterraneous, E. coli, G. pickeringii) | Algorithm: CNN with two 1D layers with 42 filters of seven units and one stride unit, a ReLU function. | 1. Can expand research by accommodating various deep learning models |
| [58] | Training dataset  F. vesca: 9708  R. Chinensis: 4674  Cross-species: 13064  Independent dataset  F. vesca: 3234  R. Chinensis: 1558  Cross species:4628 | Algorithm: CNN  Encoding techniques: DNC, TNC, BE, NCP, NCPNF, MMI Accuracy of 92% | 1. The methodology can be incorporated into the benchmark dataset of six species. |
| [59] | Benchmark dataset of three species (D. Melanogaster  C. Elegans, A. thaliana) 40,000 positive and negative samples of each species. | The neural network contained four layers and 89.3% accuracy was achieved in C.elegans, the highest compared to other species | 1. Limited use of dataset can make use of the benchmark dataset of six species |
| [60] | Expanded dataset of C.elegans 17,808 samples. | Algorithm: KNN, LR, DT, RF, SVM, NB, and three-CNN and BLTSM  Accuracy: up to 93% | 1. Better feature selection techniques such as Word2Vec can be employed. 2. Sequence length can be increased from 41 bp for further research |
| [61] | Two Benchmark dataset of six species (D. Melanogaster, C. Elegans, A. Thaliana, G. Subterraneous, E. coli, G. pickeringii) | 1st benchmark dataset model:  LR, SVM, RF, AB, GB, SGD  2ND Benchmark dataset model  CNN with attention mechanism | 1. Only Sequence information and chemical properties are included in study. 2. Information of structure and the gene expression are not included. 3. Combination of experimental prediction and validation can provide more insight. |
| [62] | Two Benchmark dataset of six species (D. Melanogaster, C. Elegans, A. Thaliana, G. Subterraneous, E. coli, G. pickeringii) | CNN and BLTSM And incorporated with attention mechanism | 1. There is no use of Automated ML/DL frameworks (PyTorch, AutoKeras). Since, they are less time consuming and are require less computational resources. |

## **2.8. Research gaps**

In the systematic literature review of 4mC site prediction, we observe that despite various machine learning and deep neural network predictors, we sense there is still potential for development, particularly in algorithm learning methodologies. The research gaps and possibly effective tactics for developing 4mC site prediction algorithms are discussed below.

1. The preliminary problem is determining what Machine Learning/Deep Learning framework to use for model training. The bulk of existing methodologies rely on manual experimental collection of the best ML/DL framework to develop the predictive model, which is a laborious and difficult procedure because there are so many Machine Learning/Deep Learning frameworks and algorithms to choose.
2. Deep Learning frameworks frequently need a significant amount of computing resources and time to train and improve the model. In this case, automatic machine learning (AutoML) programmes such as Auto-PyTorch (https://github.com/automl/Auto-PyTorch) and AutoKeras (https://autokeras.com/) are recommended for identifying the best deep neural network design.
3. However, due to a lack of experimental inquiry, the limitations of present forecasting approaches remain that only sequence information and chemical attributes are examined. When these data for 4mC sites become available, other information, like as structural data and gene expression data, should be evaluated.
4. Furthermore, such tools can aid in the model optimization process in DL model training, considerably facilitating model training and improving the resilience of learned models.

## **2.9. Research Model Development**

In this study, the dataset is collected from from METHSMRT, the benchmark dataset of six organisms (E. Coli D., Melanogastor, C. Elegans ,A. Thaliana ,G. pickeringii ,G. subterraneus) is filtered with the help of its advanced search feature. The dataset comprises equal number of positive and negative samples for each organism. One of the most important requirements of deep neural networks is that the data must be in numeric form in order for the neuron layers to process it. As a result, the next step is to convert the benchmark dataset into its numerical equivalent.

The DNNs used in this study have two types of recurrent neural networks (RNNs) based on Long Short Term Memory (LSTM) and Gated Recurrent Cells (GRCs) (GRU). In addition, Convolutional Neural Networks (CNN) are employed in this study. Chow's Pseudo k-tuple nucleotide Composition (PseKNC) is used in our research. Chou's five-step technique combines feature extraction and model training, and it is used with Deep Neural Networks to help create deep representations that can predict exact outcomes. The term "deep" refers to the overall number of layers in the model, but it may also refer to the path between the input and output nodes. Deep learning has a significant influence in the field of bioinformatics, particularly in the areas of genetic therapy and medical visualization investigation.

In the discipline of bioinformatics, deep learning technique has become the state of the art. The majority of the works employ machine-learning techniques; however, recent research suggests that deep learning models may be used. When compared to basic Machine-learning predictions, these deep learning models are significantly more efficient and produce superior outcomes. Another significant benefit of deep learning is that it eliminates the need for costly feature engineering and extraction. There is still opportunity for development in several of the key deep learning models that have been built. Our technique will ensure that these enhancements are achieved in the most efficient manner feasible.

## **2.10. Chapter Summary**

The state-of-the-art research on the research topic of DNA N4-methylcytosine prediction is reviewed in this chapter. The origins of these issues are discussed, as well as the shortcomings of present solutions. In general, computational approaches have been investigated to lower the cost of wet-lab experiments, but state-of-the-art technologies have failed to deliver sufficient results.

Existing studies are nevertheless plagued with issues. Machine learning is the most prevalent computational methodology for epigenetic mutation detection challenges, where the challenge is framed as binary classification issues for machine learning algorithms. The absence of meaningful feature representation and efficient machine learning approaches are the fundamental constraints of these technologies. Despite the fact that a continual stream of DNA 4mC sites has been discovered, the biological or regulatory role of the majority of these sites, as well as their substrates, remains largely unclear. As a result, integrating computational prediction with experimental validation will yield better results for future investigations of 4mC functions.

# **References:**

[1] D. B. Searls, “Computational biology,” *Comput. Biol. Encycl. Br.*, [Online]. Available: https://www.britannica.com/science/computational-biology.

[2] J. Gauthier, A. T. Vincent, S. J. Charette, and N. Derome, “A brief history of bioinformatics,” *Brief. Bioinform.*, vol. 20, no. 6, pp. 1981–1996, 2019, doi: 10.1093/bib/bby063.

[3] D. E. Krane and R. Micheal L, *Fundamental concepts of bioinformatics*. San Francisco : Benjamin Cummings, 2003.

[4] F. Nino, “Bioinformatics, Computational.,” *Encycl. Sci. Relig.*, 2013, doi: https://doi.org/10.1007/978-1-4020-8265-8.

[5] J. Cohen, “Computer science and bioinformatics,” *Commun. ACM*, vol. 48, no. 3, pp. 72–78, 2005, doi: 10.1145/1047671.1047672.

[6] Z. Ezziane, “Applications of artificial intelligence in bioinformatics: A review,” *Expert Syst. Appl.*, vol. 30, no. 1, pp. 2–10, 2006, doi: 10.1016/j.eswa.2005.09.042.

[7] M. J. Zaki, G. Karypis, and J. Yang, “Data mining in bioinformatics (BIOKDD),” *Algorithms Mol. Biol.*, vol. 2, no. 1, pp. 1–2, 2007, doi: 10.1186/1748-7188-2-4.

[8] M. Yuan, “Study on the Application of Data Mining in Bioinformatics,” no. Icmeit, pp. 251–255, 2016, doi: 10.2991/icmeit-16.2016.21.

[9] G. A. Pavlopoulos *et al.*, “Using graph theory to analyze biological networks,” *BioData Min.*, vol. 4, no. 1, pp. 1–27, 2011, doi: 10.1186/1756-0381-4-10.

[10] W. Huber, V. J. Carey, L. Long, S. Falcon, and R. Gentleman, “Graphs in molecular biology,” *BMC Bioinformatics*, vol. 8, no. SUPPL. 6, pp. 1–14, 2007, doi: 10.1186/1471-2105-8-S6-S8.

[11] S. N. Das, “Soft Computing Applications in BioInformatics : A Succinct Study,” no. April, pp. 1–6, 2017.

[12] S. Mitra and Y. Hayashi, “Bioinformatics with soft computing,” *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 36, no. 5, pp. 616–635, 2006, doi: 10.1109/TSMCC.2006.879384.

[13] K. J. Archer, K. Dobbin, S. Biswas, R. S. Day, D. C. Wheeler, and H. Wu, “Computer simulation, bioinformatics, and statistical analysis of cancer data and processes,” *Cancer Inform.*, vol. 14, pp. 247–251, 2015, doi: 10.4137/CIN.S32525.

[14] P. Prentice and H. Ptr, *Bioinformatics.Computing*. 2002.

[15] H. Peng, “Bioimage informatics: A new area of engineering biology,” *Bioinformatics*, vol. 24, no. 17, pp. 1827–1836, 2008, doi: 10.1093/bioinformatics/btn346.

[16] P. Cinaglia *et al.*, “Bioinformatics Solutions for Image Data Processing,” *Med. Biol. Image Anal.*, 2018, doi: 10.5772/intechopen.76459.

[17] H. Singh, “Bioinformatics: Benefits to mankind,” *Int. J. PharmTech Res.*, vol. 9, no. 4, pp. 242–248, 2016.

[18] C. Ennis, “Epigenetics 101: a beginner’s guide to explaining everything.,” *The Guardian newspaper.*, 2014.

[19] M. G. Marinus and J. Casadesus, “Roles of DNA adenine methylation in host-pathogen interactions: Mismatch repair, transcriptional regulation, and more,” *FEMS Microbiol. Rev.*, vol. 33, no. 3, pp. 488–503, 2009, doi: 10.1111/j.1574-6976.2008.00159.x.

[20] P. A. Jones and S. B. Baylin, “The fundamental role of epigenetic events in cancer,” *Nat. Rev. Genet.*, vol. 3, no. 6, pp. 415–428, 2002, doi: 10.1038/nrg816.

[21] K. D. Robertson, “DNA methylation and human disease,” *Nat. Rev. Genet.*, vol. 6, no. 8, pp. 597–610, 2005, doi: 10.1038/nrg1655.

[22] S. Kriaucionis and N. Heintz, “The nuclear DNA base 5-hydroxymethylcytosine is present in purkinje neurons and the brain,” *Science (80-. ).*, vol. 324, no. 5929, pp. 929–930, 2009, doi: 10.1126/science.1169786.

[23] H. Hayatsu and M. Shiragami, “Reaction of Bisulfite with the 5-Hydroxymethyl Group in Pyrimidines and in Phage DNAs,” *Biochemistry*, vol. 18, no. 4, pp. 632–637, 1979, doi: 10.1021/bi00571a013.

[24] S. J. Clark, A. Statham, C. Stirzaker, P. L. Molloy, and M. Frommer, “DNA methylation: Bisulphite modification and analysis,” *Nat. Protoc.*, vol. 1, no. 5, pp. 2353–2364, 2006, doi: 10.1038/nprot.2006.324.

[25] S. Tardy-Planechaud, J. Fujimoto, S. S. Lin, and L. C. Sowers, “Solid phase synthesis and restriction endonuclease cleavage of oligodeoxynucleotides containing 5-(hydroxymethyl)-cytosine,” *Nucleic Acids Res.*, vol. 25, no. 3, pp. 553–558, 1997, doi: 10.1093/nar/25.3.553.

[26] M. Tahiliani *et al.*, “Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1,” *Science (80-. ).*, vol. 324, no. 5929, pp. 930–935, 2009, doi: 10.1126/science.1170116.

[27] Y. Li, C. Huang, L. Ding, Z. Li, Y. Pan, and X. Gao, “Deep learning in bioinformatics: Introduction, application, and perspective in the big data era,” *Methods*, vol. 166, no. March, pp. 4–21, 2019, doi: 10.1016/j.ymeth.2019.04.008.

[28] K. Lan, D. tong Wang, S. Fong, L. sheng Liu, K. K. L. Wong, and N. Dey, “A Survey of Data Mining and Deep Learning in Bioinformatics,” *J. Med. Syst.*, vol. 42, no. 8, 2018, doi: 10.1007/s10916-018-1003-9.

[29] B. Tang, Z. Pan, K. Yin, and A. Khateeb, “Recent advances of deep learning in bioinformatics and computational biology,” *Front. Genet.*, vol. 10, no. MAR, 2019, doi: 10.3389/fgene.2019.00214.

[30] M. Hapudeniya, “Artificial Neural Networks in Bioinformatics,” *Sri Lanka J. Bio-Medical Informatics*, vol. 1, no. 2, p. 104, 2010, doi: 10.4038/sljbmi.v1i2.1719.

[31] L. J. Lancashire, C. Lemetre, and G. R. Ball, “An introduction to artificial neural networks in bioinformatics - Application to complex microarray and mass spectrometry datasets in cancer studies,” *Brief. Bioinform.*, vol. 10, no. 3, pp. 315–329, 2009, doi: 10.1093/bib/bbp012.

[32] S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” *Proc. 2017 Int. Conf. Eng. Technol. ICET 2017*, vol. 2018-January, pp. 1–6, 2018, doi: 10.1109/ICEngTechnol.2017.8308186.

[33] J. Wu, “Introduction to Convolutional Neural Networks,” *National Key Lab for Novel Software Technology Nanjing University, China*, 2017. .

[34] M. Gurucharan, “Basic CNN Architecture: Explaining 5 Layers of Convolutional Neural Network,” *upGrad*, 2020. https://www.upgrad.com/blog/basic-cnn-architecture/.

[35] S. J. Fong and R. C. Millham, *Bio-inspired Algorithms for Data Streaming and Visualization, Big Data Management, and Fog Computing*. 2021.

[36] A. Sherstinsky, “Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network,” *Phys. D Nonlinear Phenom.*, vol. 404, p. 132306, 2020, doi: 10.1016/j.physd.2019.132306.

[37] B. A. Flusberg *et al.*, “Direct detection of DNA methylation during single-molecule, real-time sequencing,” *Nat. Methods*, vol. 7, no. 6, pp. 461–465, 2010, doi: 10.1038/nmeth.1459.

[38] T. A. Clark *et al.*, “Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing,” *Nucleic Acids Res.*, vol. 40, no. 4, 2012, doi: 10.1093/nar/gkr1146.

[39] C. Grunau, E. Renault, A. Rosenthal, and G. Roizes, “MethDB - A public database for DNA methylation data,” *Nucleic Acids Res.*, vol. 29, no. 1, pp. 270–274, 2001, doi: 10.1093/nar/29.1.270.

[40] M. Hackenberg, G. Barturen, and J. L. Oliver, “NGSmethDB: A database for next-generation sequencing single-cytosine- resolution DNAmethylation data,” *Nucleic Acids Res.*, vol. 39, no. SUPPL. 1, pp. 75–79, 2011, doi: 10.1093/nar/gkq942.

[41] D. Zou, S. Sun, R. Li, J. Liu, J. Zhang, and Z. Zhang, “MethBank: A database integrating next-generation sequencing single-base-resolution DNA methylation programming data,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D54–D58, 2015, doi: 10.1093/nar/gku920.

[42] X. He *et al.*, “MethyCancer: The database of human DNA methylation and cancer,” *Nucleic Acids Res.*, vol. 36, no. SUPPL. 1, pp. 836–841, 2008, doi: 10.1093/nar/gkm730.

[43] S. J. Baek, S. Yang, T. W. Kang, S. M. Park, Y. S. Kim, and S. Y. Kim, “MENT: Methylation and expression database of normal and tumor tissues,” *Gene*, vol. 518, no. 1, pp. 194–200, 2013, doi: 10.1016/j.gene.2012.11.032.

[44] W. Y. Huang *et al.*, “MethHC: A database of DNA methylation and gene expression in human cancer,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D856–D861, 2015, doi: 10.1093/nar/gku1151.

[45] M. Ongenaert, L. Van Neste, T. De Meyer, G. Menschaert, S. Bekaert, and W. Van Criekinge, “PubMeth: A cancer methylation database combining text-mining and expert annotation,” *Nucleic Acids Res.*, vol. 36, no. SUPPL. 1, pp. 842–846, 2008, doi: 10.1093/nar/gkm788.

[46] P. Ye, Y. Luan, K. Chen, Y. Liu, C. Xiao, and Z. Xie, “MethSMRT: An integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D85–D89, 2017, doi: 10.1093/nar/gkw950.

[47] W. Chen, H. Yang, P. Feng, H. Ding, and H. Lin, “IDNA4mC: Identifying DNA N 4 -methylcytosine sites based on nucleotide chemical properties,” *Bioinformatics*, vol. 33, no. 22, pp. 3518–3523, 2017, doi: 10.1093/bioinformatics/btx479.

[48] W. He, C. Jia, and Q. Zou, “4mCPred: Machine learning methods for DNA N 4 -methylcytosine sites prediction,” *Bioinformatics*, vol. 35, no. 4, pp. 593–601, 2019, doi: 10.1093/bioinformatics/bty668.

[49] K. C. Chou, “Some remarks on protein attribute prediction and pseudo amino acid composition,” *J. Theor. Biol.*, vol. 273, no. 1, pp. 236–247, 2011, doi: 10.1016/j.jtbi.2010.12.024.

[50] L. Wei, S. Luan, L. A. E. Nagai, R. Su, and Q. Zou, “Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species,” *Bioinformatics*, vol. 35, no. 8, pp. 1326–1333, 2019, doi: 10.1093/bioinformatics/bty824.

[51] M. M. Hasan, B. Manavalan, M. S. Khatun, and H. Kurata, “i4mC-ROSE, a bioinformatics tool for the identification of DNA N4-methylcytosine sites in the Rosaceae genome,” *Int. J. Biol. Macromol.*, vol. 157, no. xxxx, pp. 752–758, 2020, doi: 10.1016/j.ijbiomac.2019.12.009.

[52] L. Wei *et al.*, “Iterative feature representations improve N4-methylcytosine site prediction,” *Bioinformatics*, vol. 35, no. 23, pp. 4930–4937, 2019, doi: 10.1093/bioinformatics/btz408.

[53] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, “Meta-4mCpred: A Sequence-Based Meta-Predictor for Accurate DNA 4mC Site Prediction Using Effective Feature Representation,” *Mol. Ther. - Nucleic Acids*, vol. 16, no. June, pp. 733–744, 2019, doi: 10.1016/j.omtn.2019.04.019.

[54] B. Manavalan, S. Basith, T. H. Shin, D. Y. Lee, L. Wei, and G. Lee, “4mCpred-EL: An Ensemble Learning Framework for Identification of DNA N4-methylcytosine Sites in the Mouse Genome,” *Cells*, vol. 8, no. 11, pp. 1–14, 2019, doi: 10.3390/cells8111332.

[55] J. Yang, K. Lang, G. Zhang, X. Fan, Y. Chen, and C. Pian, “SOMM4mC: a second-order Markov model for DNA N4-methylcytosine site prediction in six species,” *Bioinformatics*, vol. 36, no. 14, pp. 4103–4105, 2020, doi: 10.1093/bioinformatics/btaa507.

[56] Q. Tang *et al.*, “DNA4mC-LIP: a linear integration method to identify N4-methylcytosine site in multiple species,” *Bioinformatics*, vol. 36, no. 11, pp. 3327–3335, Jun. 2020, doi: 10.1093/bioinformatics/btaa143.

[57] F. Zeng, G. Fang, and L. Yao, “A Deep Neural Network for Identifying DNA N4-Methylcytosine Sites,” *Front. Genet.*, vol. 11, 2020, doi: 10.3389/fgene.2020.00209.

[58] J. Khanal, I. Nazari, H. Tayara, and K. T. Chong, “4mCCNN: Identification of N4-Methylcytosine Sites in Prokaryotes Using Convolutional Neural Network,” *IEEE Access*, vol. 7, pp. 145455–145461, 2019, doi: 10.1109/ACCESS.2019.2943169.

[59] A. Wahab, O. Mahmoudi, J. Kim, and K. T. Chong, “DNC4mC-Deep: Identification and Analysis of DNA N4-Methylcytosine Sites Based on Different Encoding Schemes By Using Deep Learning,” *Cells*, vol. 9, no. 8, 2020, doi: 10.3390/cells9081756.

[60] R. Zeng and M. Liao, “Developing a Multi-Layer Deep Learning Based Predictive Model to Identify DNA N4-Methylcytosine Modifications,” *Front. Bioeng. Biotechnol.*, vol. 8, no. April, 2020, doi: 10.3389/fbioe.2020.00274.

[61] H. Xu, P. Jia, and Z. Zhao, “Deep4mC: systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning,” *Brief. Bioinform.*, vol. 22, no. 3, pp. 1–13, 2021, doi: 10.1093/bib/bbaa099.

[62] Q. Liu *et al.*, “DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites,” *Brief. Bioinform.*, vol. 22, no. 3, pp. 1–14, 2021, doi: 10.1093/bib/bbaa124.