HARNESSING DATA SCIENCE AND BIG DATA IN BIOINFORMATICS: A COMPREHENSIVE REVIEW OF COMPUTATIONAL TECHNIQUES AND APPLICATIONS

Ayan Das[1], Mousumi Ghosh[2], Subhadip Das[3], Aditi Karmakar[4], Subrata Kumar Majumdar[5] and Kousik Roy[3]*

[1]Department of Computer Science and Technology, Institute of Engineering and Management (IEM) Newtown, Kolkata – 700160

[2] Institute of Engineering and Management (IEM) Newtown, Kolkata – 700160

[3]Department of Computer Science and Engineering, Bengal College of Engineering and Technology, Durgapur, West Bengal

[4]Department of Computer Science Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, Odisha

[5]Department of Mechanical Engineering, Sanaka Educational Trust's Group of Institutions, Durgapur, West Bengal

*Corresponding Author Email Id: kousikroy002@gmail.com

ABSTRACT

Advances in high-throughput technologies have led to a surge in biological data, making bioinformatics a key player in data-driven research. This review explores the intersection of Data Science and Big Data with bioinformatics, highlighting machine learning, statistical modelling and scalable analytics tools like Hadoop and Spark. Key applications include genomics, disease classification and biomarker identification. It also examines public databases (e.g., TCGA, ENCODE), data integration challenges and ethical considerations. Emerging trends such as deep learning, federated learning, and multi-omics integration are discussed. The review provides insights into current practices and outlines future directions in computational biology and precision medicine.

Keywords: Bioinformatics, Data Science, Big Data Analytics, Machine Learning, Deep Learning, Statistical Modelling

1. INTRODUCTION

Modern bioinformatics is being reshaped by the convergence of large biological datasets and advances in deep generative modelling [1],[2]. Diffusion models now outperform traditional approaches like variational auto encoders and GANs for tasks such as molecule and protein structure generation, omics data synthesis and noise reduction in biological imaging [3],[4]. These models are fundamentally grounded in score-based mathematical constructs and SDEs, enabling robust sampling and high-fidelity data transformation for complex biological systems [5],[6].

2. TECHNICAL FOUNDATIONS

2.1 Denoising Score Matching and Diffusion Models

Diffusion models learn the data distribution by simulating a forward process that gradually corrupts data with noise and a reverse process that learns to denoise step by step [6],[7]. DSM trains a neural network to predict the gradient of the log-density (the "score") of noisy data [8]. Given data x_0, the forward process adds Gaussian noise: $q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \boldsymbol{I})$,

where β_t controls noise magnitude at each time step. The reverse process parameterized by ϑ approximates: