Chapter 2
An Overview on Bioinformatics

ABSTRACT

This chapter presents a thorough background and deep literature review of the current topic of study. It also presents and defines the key concepts utilised throughout this investigation. It consists of ten sections: (1) a background on bioinformatics, (2) a discussion of colon cancer, (3) an overview of the microarray technology that is used to extract the dataset, (4) an overview of the colon cancer dataset, (5) a review of the most prevalent algorithms employed for gene selection and cancer classification, (6) a presentation of related works from the literature, (7) identification of feature selection approaches and procedures, (8) an investigation of the ML concept, (9) a review of algorithm efficiency and time complexity analysis, and (10) identification of current problems in the research area.

1. INTRODUCTION

This chapter presents a thorough background and deep literature review of the current topic of study. It also presents and defines the key concepts utilised throughout this investigation. It consists of ten sections: (1) a background on bioinformatics; (2) a discussion of colon cancer; (3) an overview of the microarray technology that is used to extract the dataset; (4) an overview of the colon cancer dataset; (5) a review of the most prevalent algorithms employed for gene selection and cancer classification; (6) a presentation of related works from the literature; (7) identification of feature selection approaches and procedures; (8) an investigation of the ML concept; (9) a review of algorithm efficiency and time complexity analysis; and (10) identification of current problems in the research area.

2. BIOINFORMATICS: AN OVERVIEW OF CANCER RESEARCH

Bioinformatics is the integration of the fields of biology, computer science, statistics, and mathematics (Al-Rajab & Lu, 2012), with each field playing a significant role in gathering, forming, analysing, and digitising genetic data. Moreover, it aids the efficient categorisation and storage of data (Al-Rajab & Lu, 2012; Bayat, 2002; Cohen, 2005; Jawdat, 2006; Jena et al., 2009; Ng & Wong, 2004).

DOI: 10.4018/978-1-7998-7316-7.ch002

Bioinformatics was first introduced in 1979 by Paulien Hogeweg to study the informatic procedures of biological systems (Raza, 2012). This section elaborates bioinformatics using a variety of scientific papers to establish the basis for the current research, to identify core bioinformatics applications, to present the data structure and the central databases employed, to provide an overview of the most popular algorithms applied to this field, and to advocate for the employment of bioinformatics in the area of cancer research.

2.1 Background of Methodologies

Several scientific papers, books, articles, and websites provide information regarding bioinformatics. However, these resources do not present a unified, official, or integrated description for bioinformatics as a science. Most works simply present basic descriptions. Therefore, a deep search was conducted to gather as much knowledge as possible to understand the field and connect its significance to cancer research.

Jawdat (2006) described bioinformatics as a process using specific methods, algorithms, and computer software to store and interpret biological data. It is the design, implementation, and employment of computer tools to produce, store, interpret, access, and analyse molecular biology data. (Raut, Sathe, & Raut, 2010) stated that bioinformatics is fundamentally the study to design, arrange, grasp and explore useful information linked to a large amount biotic data. The term, 'bio' signifies the molecular biology aspect, and 'informatics' reflects the information technology used to control, interpret, and employ vast amounts of genetic data. In (Ng & Wong, 2004), the authors claimed that bioinformatics manages, organises and analyses genetic data, which is the raw, basic essence of any organism. Bioinformatics connects computer science, biology and mathematics to address the many computational challenges of modern medical research. Chavan (2008) maintained that genetic data contains a wide range of information concerning genetic sequences regulating diversity, evolutionary changes and alterations of proteins. Bioinformatics is a product of the need to systematically filter data for classification and indexing. Thus, it can be defined as the field that combines a diversity of sciences for this purpose (Al-Rajab & Lu, 2012). It is the discipline of controlling, interpreting and analysing a massive amount of genetic data using progressive computing methods and techniques. Furthermore, other authors claimed that the difference between bioinformatics and computational biology cannot be simply deduced (Ackovska & Madevska-Bogdanova, 2005). Both topics involve several scientific disciplines, including molecular biology, computer science, mathematics, statistics, physics, biochemistry and genetics (Al-Rajab & Lu, 2012, Guo et al, 2021). Zadeh (2006) defined bioinformatics as a modern field developed from the domains of biology, computer science and biochemistry. Bioinformatics is a multidisciplinary study and a fast-growing topic evolved from the domains of biology, computer science and chemistry. Additionally, Kasabov (2004) claimed that bioinformatics included the development of information science applications needed to analyse, demonstrate and discover knowledge of biological activities in living organisms. Moreover, in (Fulekar & Sharma, 2008), the authors focused on the mixture of information technology and biology, insisting that bioinformatics concentrated on the molecular cell level of biotechnology. Fenstermacher (2005) defined bioinformatics as a multidimensional topic joining many specialised disciplines, such as computational biology, mathematics, statistics, molecular chemistry and biology. Additionally, Nair (2007) stated that bioinformatics answers biologists' questions about the ambiguities of life via the application of computer science and associated technologies. Other authors stated that bioinformatics is a modern and fast-developing topic that combines molecular biology, biochemistry, artificial intelligence (AI), databases, pattern recognition and computer science algorithms (Al-Rajab & Lu, 2012; Doom et al.,