IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 2, 2022, accepted March 16, 2022, date of publication March 21, 2022, date of current version March 25, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3160795

Exploratory Analysis of Topic Interests and Their Evolution in Bioinformatics Research Using Semantic Text Mining and Probabilistic Topic Modeling

FATIH GURCAN[1], **AND NERGIZ ERCIL CAGILTAY**[2]
[1]Department of Computer Engineering, Faculty of Engineering, Karadeniz Technical University, 61080 Trabzon, Turkey
[2]Department of Software Engineering, Faculty of Engineering, Atilim University, 06830 Ankara, Turkey

Corresponding author: Fatih Gurcan (fgurcan@ktu.edu.tr)

ABSTRACT Bioinformatics, which has developed rapidly in recent years with the collaborative contributions of the fields of biology and informatics, provides a deeper perspective on the analysis and understanding of complex biological data. In this regard, bioinformatics has an interdisciplinary background and a rich literature in terms of domain-specific studies. Providing a holistic picture of bioinformatics research by analyzing the major topics and their trends and developmental stages is critical for an understanding of the field. From this perspective, this study aimed to analyze the last 50 years of bioinformatics studies (a total of 71,490 articles) by using an automated text-mining methodology based on probabilistic topic modeling to reveal the main topics, trends, and the evolution of the field. As a result, 24 major topics that reflect the focuses and trends of the field were identified. Based on the discovered topics and their temporal tendencies from 1970 until 2020, the developmental periods of the field were divided into seven phases, from the ''newborn'' to the ''wisdom'' stages. Moreover, the findings indicated a recent increase in the popularity of the topics ''Statistical Estimation'', ''Data Analysis Tools'', ''Genomic Data'', ''Gene Expression'', and ''Prediction''. The results of the study revealed that, in bioinformatics studies, interest in innovative computing and data analysis methods based on artificial intelligence and machine learning has gradually increased, thereby marking a significant improvement in contemporary analysis tools and techniques based on prediction.

INDEX TERMS Bioinformatics corpus, probabilistic topic modeling, textual content analysis, scientometric analysis, bioinformatics topics and trends.

I. INTRODUCTION

With the contributions of biology and computer science, bioinformatics is becoming one of the critical fields providing a deeper understanding of biological data. Bioinformatics brings together biology, computer science, mathematics, statistics, and information technologies under the same umbrella [1], [2]. Accordingly, the field of bioinformatics is considered an interdisciplinary field that develops methods and tools for the analysis, interpretation, and synthesis of biological data containing large and complex datasets [2], [3]. Because of this interdisciplinary nature of the field, rapidly developing technologies also influence the studies conducted under bioinformatics. In this context, bioinformatics methods and tools provide methodological perspectives that enable the analysis, comparison, and interpretation of genetic and genomic data [4]. Bioinformatics also plays a role in the analysis of gene and protein expression and in the development of biological gene ontologies [3]. Bioinformatics is generally considered to be a synonym for computational biology [5]–[7]. On the other hand, it is a different field of science from biological computing [5].

Bioinformatics was coined as a term by Hogeweg and Hesper in [8] and has become an important part of many fields of biology today [4]. The bioinformatics field began with an explosive growth in the 1970s that accelerated in the

The associate editor coordinating the review of this manuscript and approving it for publication was Rajesh Kumar.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

1990s, largely as a result of the Human Genome Project and significant advances in DNA, RNA, and protein sequencing technologies [1], [6]. In this context, bioinformatics studies have been carried out for over fifty years, and the number of studies in this field has increased exponentially [2], [4]. It has been reported that around 34% of the top-cited science studies published between 1994 and 2013 are in the field of bioinformatics [9]. Major research and application efforts in the field include focal subjects, such as the alignment of protein structure and protein sequence, genome assembly, drug design and discovery, analysis of genomic data, prediction of gene expression, genome annotation, building biological networks, protein-protein interaction, transcription, species population and evolution, tumor cells, statistical estimation, algorithms, ontology-based information retrieval, and development of analysis and software tools [1], [3], [4], [10], [11].

In that respect, providing a holistic view of the field is critical and important for planning future directions and for a better understanding of the explored topics and trends and their developmental stages [12]. The analysis of the research corpus of bioinformatics from a deeper and broader perspective is very important for the understanding of the hidden semantic patterns (topics) and trends of the field and thus, for better planning, guiding, and supporting of future research. In the literature, there have been several attempts to provide such a picture of bioinformatics studies [13]–[16]. Despite these valuable efforts, the use of topic modeling approaches to reveal the research landscape of the bioinformatics field is still limited. This study aimed to fill this gap in the literature through a semantic content analysis based on a topic modeling approach implemented on 71,490 articles published between 1970 and 2020 in the top 20 publication sources specific to the bioinformatics field. The bibliometric characteristics, research topics, temporal trends, correlations, developmental stages, and future directions of bioinformatics were analyzed and are presented here. More specifically, by considering the last 50 years of the bioinformatics field, the methodology of this study was designed to seek answers to the following research questions (RQ):

RQ1. What are the bibliometric characteristics of bioinformatics research?

RQ2. What are the emerging topics of bioinformatics research?

RQ3. How have the topics of bioinformatics changed over the last 50 years?

RQ4. What are the developmental stages of bioinformatics?

RQ5. What is the future trend of bioinformatics?

II. RELATED WORK

In the literature, several studies have attempted to better understand the trends of bioinformatics research and a number of studies have been conducted to better understand the big data in bioinformatics [17], [18]. For example, in a recent study, da Silva *et al.* [19] listed twenty trend topics by analyzing a set of articles about the use in bioinformatics of big data that had been collected from three scientific bases, i.e., Scopus, ACM, and Web of Science. However, as these studies considered only the big-data perspective of the field, they present a limited picture of bioinformatics studies in general.

There are also very few studies examining the entire perspective of the field. For example, Wu *et al.* [14] applied text-mining algorithms to analyze 678 abstracts (1986-1995) and 4,961 full articles (1996-2008) from Bioinformatics Journal. They discovered 30 topics in the field by deducing the growth trends of genome analysis and a slow decline of historically popular topics. However, as their corpus was created through a specific journal (Bioinformatics Journal), their results demonstrated a narrow representation of bioinformatics literature. Similarly, Song *et al.* [20] analyzed 13,508 studies published in 33 bioinformatics-related conferences indexed in DBLP from 2000 to 2011. Although they showed some trends in the field, their scope was also based on articles published in the conferences; therefore, they also presented a limited picture of bioinformatics studies. Heo *et al.* [16] analyzed 241,569 articles published between 1996 and 2015 in 46 PubMed-listed journals using LDA and concluded that by using applied mathematics, information science, statistics, computer science, chemistry, and biochemistry, bioinformatics studies mainly tackled biological problems at the molecular level. However, an analysis of the list of these journals reveals that the scope of some of them, such as PLOS Genetics, PLOS Biology, Trends in Genetics, and Journal of Computational Neuroscience, covers an enhanced set of topics other than bioinformatics. Although they analyzed 241,569 articles published in these journals, there is a high possibility that their corpus included studies not directly related to the field of bioinformatics. Similarly, Patra and Mishra [21] conducted a study by creating a corpus from PubMed. However, as bioinformatics is an interdisciplinary field, using only PubMed as a corpus source has some limitations. There is a possibility that some bioinformatics studies might have been published in technical conferences and journals, which are not indexed in PubMed. Hence, examining the publications indexed only in PubMed also limited their corpus of bioinformatics studies.

In another study, Ouzounis [22] analyzed the use of the term ''bioinformatics'' in Google trends and classified its trends as ''Infancy'' (1996–2001), ''Adolescence'' (2002–2006), and ''Adulthood'' (2007–2011). The author also concluded that computational biology would emerge as a distinct discipline in the near future. Hahn *et al.* [23] conducted an LDA-based topic model analysis by using the term ''bioinformatics'' to perform a search in the Scopus database that extended to 85,106 articles published between 1998 and 2016. They listed 25 keywords and conducted a temporal analysis of these keywords. They reported that keywords such as ''big data'', ''next-generation sequencing'', and ''cancer'' were increasing in popularity, whereas after 2000, the keyword ''drug discovery'' reached a plateau [23]. Their study conducted a keyword analysis of the publications. In addition to author keywords, titles and abstracts of articles are also

important to better understand the scope of a study; therefore, their study could be considered limited in its representation of the general picture of bioinformatics literature.

Recently, Youssef and Rich [7] conducted topic modeling on 143,545 bioinformatics studies published between 1987 and 2018 and by applying LDA analysis to this corpus, identified eleven main categories. Their results indicated that cancer biology and clinical informatics were the most popular research topics. However, as they searched only for the terms ''bioinformatics'' and ''computational biology'' in PubMed, there is a possibility that they might have missed some important studies related to bioinformatics that did not explicitly use these keywords. Additionally, they did not report the volume of each topic, which prevents us from seeing the status of each topic in the field. A recent study conducted by Hahn *et al.* [23] is also limited to a keyword analysis, omitting the titles and abstracts of the analyzed studies.

For the corpus creation, some topic modeling attempts in the field of bioinformatics are based on a keyword search (e.g., [15]), or are specific to a country (e.g., [24]) or a productivity analysis [20]. Most of these studies have limitations (e.g., [25]) in their corpus creation method. Because of the importance of aiding biomedical science researchers interested in analyzing research trends and developing insights, Kavvadias *et al.* [26] developed an easy-to-use open-source tool for utilizing the huge amount of research trends. Their approach is especially valuable for standardizing and analyzing the studies in the field. The corpus and text mining algorithms used for these analyses are critical because they affect the results [27], [28]. For this reason, a standardized corpus creation approach for these analyses is necessary.

As seen from the earlier studies, bioinformatics is a long-standing and interdisciplinary field. Although there have been several attempts to uncover the trends of the studies conducted in this field, they either offer a limited perspective of the field or analyze studies indirectly related to bioinformatics. Therefore, excluding some related work, their corpus is limited. In order to fill this gap in the literature, this study applied a systematical approach to corpus creation and analysis of bioinformatics studies by considering the authors, titles, and abstracts of the studies in order to provide a broader picture of bioinformatics literature and its trends. From this perspective, we believe that this current study provides a very valuable contribution to bioinformatics literature.

III. METHOD

The method of the study is summarized below by describing the corpus creation methodology, the data processing methodology and the implementation of the topic modeling on the created corpus.

A. DATA COLLECTION–CREATING THE CORPUS

For topic modeling studies based on textual content analysis, the creation of an empirical corpus is usually one of the most critical steps [29], [30]. The selection of the approach to be applied in the corpus creation procedure usually has a direct influence on the results [28], [29]. For this reason, an objective approach was followed for corpus creation in this study. In this regard, firstly, with the aim of creating an experimental corpus of bioinformatics, the top 20 publication sources categorized and indexed by Google Scholar Metrics for bioinformatics and computational biology were taken into account as the main data source. These sources included the 17 journals and three conferences with the highest scores according to the H5-index. Among the studies published in these sources, only research articles, conference papers, and review articles in English were included in the corpus. Other studies like book reviews or editorial materials were excluded. The number of articles included in this corpus was also checked and verified from PubMed, Scopus, and Web of Science. As PubMed does not index conference papers, these were checked from their original proceedings [31]. Afterwards, using the Scopus bibliometric database, which indexes all of the articles in these journals and conferences, the data including the title, abstract, and author keywords of each article were downloaded and saved in the experimental database [32]. As a result, an empirical corpus of bioinformatics was created that included 71,490 articles (62,784 papers, 1,194 reviews, and 7,512 conference papers) published in these sources between 1970 and 2020. The distribution of the articles according to these sources and the H5-index scores of these sources are given in Table 1.

B. DATA PREPROCESSING

Data preprocessing is a fundamental task that directly affects the success of the analysis, especially for text mining studies [27], [29], [33]. In this empirical analysis, for the preprocessing procedure required for the topic modeling process to be successfully applied to the bioinformatics corpus containing a large amount of textual data, the following sequential steps were performed. Initially, all textual contents in the dataset were converted to lowercase. Web links, tags, publisher information, numeric expressions, punctuation marks, and symbols in the dataset were then deleted. Word tokenization was implemented to characterize textual contents as words [34]–[36]. In the next step, all English stop words (how, what, and, is, or, the, a, an, for, etc.) that are often used in texts and that do not make any sense alone were deleted [34]–[36]. Likewise, generic words (e.g., literature, article, paper, research, study, and copyright) that were frequently observed in the articles but did not contribute to the creation of semantically-consistent topics were also filtered out. In this way, it was ensured that only the words with a single meaning were included in the dataset. Using the Snowball stemming algorithm [37], the remaining words in the dataset were reduced to their stems (e.g., the words ''interaction'', ''interactions'', ''interacting'', ''interacts'', and ''interacted'' were all represented as ''interact'') [27]. As a final point, each of the 71,490 articles in the dataset was transformed to a word vector using the ''bag of words'' approach to enable a numerical representation of the words in the corpus [28], [29], [34], [35]. Accordingly, a document-term

TABLE 1. Publication sources.

#	Publication Source	H5-I	N	%
1	Bioinformatics	125	14,788	19.80
2	PLOS Computational Biology	84	7,139	9.56
3	Briefings in Bioinformatics	65	1,377	1.84
4	BMC Bioinformatics	63	10,009	13.40
5	GigaScience	51	647	0.87
6	Journal of Theoretical Biology	45	14,043	18.80
7	Database: The Journal of Biological Databases & Curation	43	376	0.50
8	IEEE/ACM Transactions on Computational Biology and Bioinformatics	41	1,867	2.50
9	Journal of Cheminformatics	41	697	0.93
10	Genomics, Proteomics & Bioinformatics	36	552	0.74
11	BMC Systems Biology	31	1,922	2.57
12	Mathematical Biosciences	29	4,395	5.88
13	Journal of Mathematical Biology	29	3,032	4.06
14	IEEE International Conference on Bioinformatics and Biomedicine	25	3,918	5.25
15	Computational Biology and Chemistry	25	1,606	2.15
16	Pacific Symposium on Biocomputing	25	1,264	1.69
17	Journal of Biomedical Semantics	24	439	0.59
18	Journal of Computational Biology	23	2,086	2.79
19	Molecular Informatics	23	773	1.03
20	International Conference on Research in Computational Molecular Biology	23	560	0.75
	Total		**71,490**	**100**

matrix (DTM) representing the entire corpus and providing the matrix form required for topic modeling analysis was created by combining these vectors [34], [36], [38], [39].

C. FITTING AND IMPLEMENTATION OF TOPIC MODELING

Topic modeling is an unsupervised machine learning technique used to automatically discover hidden semantic structures called "topics" in a specific corpus [28], [34]–[36], [40]. Many different topic modeling algorithms are available for text mining and natural language processing research, such as Latent Dirichlet Allocation (LDA), Hierarchical Latent Dirichlet Allocation (HLDA), Hierarchical Dirichlet Process (HDP), Non-Negative Matrix Factorization (NMF), Dirichlet Multinomial Regression (DMR), Dynamic Topic Model (DTM) and Correlated Topic Model (CTM) [41], [42]. In some of these algorithms (NMF, CTM, and DMR), the approach that calculates a commonly accepted consistency score used to estimate the optimal number of topics has unfortunately been observed to be very limited [42]. HDP and HLDA algorithms, which automatically identify the optimal number of topics, are also recently proposed topic models [42]. Among all topic modeling algorithms, LDA constitutes the basic theory of topic modeling and is widely used. Therefore, LDA is frequently preferred in many researches and applications. In this context, we applied different models of HDP, HLDA, and LDA algorithms to the empirical corpus of bioinformatics in order to clearly compare HDP and HLDA topic modeling algorithms, which identify the number of topics in an automated way, with LDA.

As a result, HDP and HLDA algorithms need many prior parameters to identify the optimal number of the topics. For this reason, it has been observed that even very small changes in the values of these parameters significantly change the results of the analysis [41]. The optimal estimation of these parameters is an important constraint for both of these algorithms [41]. Besides, the outputs of the topic modeling analysis by HLDA were far from our expectation. Despite the most stringent parameters, huge numbers of topics (ranging from 43 to 331) were obtained in most analyzes with HLDA. In contrast, using HDP, we obtained a smaller number of topics (ranging from 13 to 29) compared to HLDA. However, the semantic consistency of these topics obtained with HDP was interpreted as insufficient by two experts in the field of bioinformatics.

Consequently, in this study, these two algorithms (HLDA and HDP) did not reveal satisfactory results in detecting the detail level and semantic consistency of the topics in the corpus. On the other hand, in LDA model, deciding to increase or decrease the number of topics by manually examining the topics discovered in the previous analysis and repeating the analysis with this iterative process provided a more effective solution to the ideal estimation of the number of topics and their semantic consistency. Moreover, LDA algorithm provides many efficient methods for calculating the coherence score for the estimation of the optimal number of topics and is widely applied in many fields [34], [43]. For these reasons, LDA is a highly preferred and accepted algorithm for the semantic content analysis of large textual corpora because of the systematic approaches it provides [28], [30], [39].

Accordingly, in this study, Latent Dirichlet Allocation (LDA) [34], [38], a probabilistic model for topic modeling-based content analysis, was used for semantic analysis of the empirical corpus of bioinformatics literature. The LDA is a generative model based on Bayesian inference that provides an effective approach to analyzing large collections of textual documents in a systematic manner; therefore, it is extensively used for topic modeling analysis [34], [38], [43], [44]. The main intuition of LDA is based on the assumption

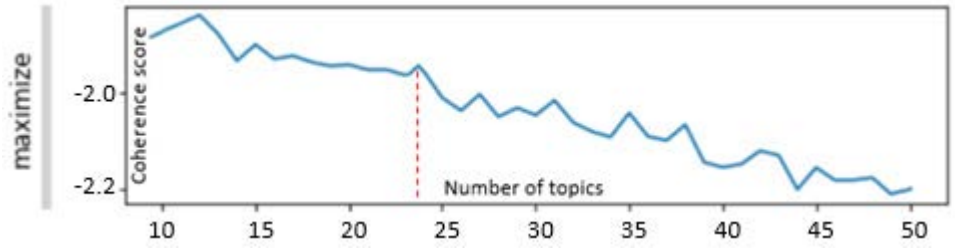

FIGURE 1. Choosing optimal model with coherence scores.

that each document in a dataset covers more than one topic. The LDA calculates the per-document topic distribution, per-topic word distribution, and per-document topic and word assignments using an iterative process based on Dirichlet distribution [34]–[36], [38].

With the aim of fitting and applying the LDA-based topic modeling procedure to the corpus of bioinformatics, this study used the tmtoolkit package [34]–[36], an extensive toolkit developed in Python for text mining and topic modeling. With the intention of fitting the LDA model to the empirical corpus, the values of prior parameters enabling the optimization of the model were initially selected [34]–[36]. The prior parameters of α, which sets the topic distribution in the documents, and β, which sets the word distribution in the topics, were used as $\alpha = 0.1$ and $\beta = 0.01$, as the recommended values for the topic modeling of short texts [29], [40].

Afterwards, in order to empirically select the ideal number of topics (K), the LDA model was employed with different K values ranging between 10 and 50 [34]–[36]. During this process, a coherence metric (C_V) was calculated for each trial conducted with each K value using the semantic coherence approach[45], as shown in Fig. 1. Considering the important breaking points (e.g. 13, 24, 27, 32 values of K) in Fig. 1, the clarity and semantic consistency of the discovered topics at these points were evaluated. Within these breaking points, after the first breaks (K = 13 and 15), a maximum coherence score ($C_V = -1.9542396$) reflecting optimal topic-word distributions and semantic consistency for each topic was achieved with the topic number of K = 24 [34], [36], [45].

The meaningfulness and consistency of these topics described by the representative keywords were also evaluated independently by two experts studying in the fields of molecular biology and biochemistry. After examining the consistency of the topics, the label of each topic was identified and assigned by the two field experts, taking into account the descriptive keywords of the topics [29].

In addition, the percentage rate per document of each topic, the distribution of the words in each topic, and the distribution rate of the topics over the entire corpus were calculated [34]–[36]. At the end of this process, the 15 representative keywords with the highest frequency were identified for each of the 24 topics. Consequently, these 24 topics discovered by LDA considering the coherence metric C_V were used in all subsequent analyses.

IV. RESULTS

The results of the study are presented under five main sections, each of which attempts to answer the above-mentioned research questions. Accordingly, the corpus was firstly analyzed to descriptively present the bibliographic characteristics of the bioinformatics studies included in the corpus, and then topic modeling analysis was applied to discover the emerging topics of bioinformatics research. Next, the study sought to reveal the change in the topics of bioinformatics over the last 50 years. Finally, the developmental stages and future trends of bioinformatics research were explored.

A. DESCRIPTIVE ANALYSIS

In order to answer the first research question (RQ1), a descriptive analysis was conducted to reveal the publication sources, the number of articles included in the corpus of the study, and their subject areas. The list of publication sources and their H5-index (H5-I) values are given in Table 1. Additionally, the table presents the number of articles from each publication source (N) and their percentages (%) in the entire corpus. Thus, among all the publication sources, Bioinformatics published the highest number of articles (19.80%).

TABLE 2. Total number of articles in 5-year increments.

Period	N	%
1971-1975	1,398	1.96
1976-1980	1,913	2.68
1981-1985	2,027	2.84
1986-1990	1,857	2.60
1991-1995	2,196	3.07
1996-2000	2,771	3.88
2001-2005	6,235	8.72
2006-2010	14,218	19.89
2011-2015	18,538	25.93
2016-2020	20,337	28.45
Total	71,490	100.00

Table 2 presents the total number of publications (N) and their percentages (%) in 5-year increments. Notably, after 2006, there was a significant increase (from 8.72% to 19.80%) in the number of articles in the field of bioinformatics.

Table 3 presents the number of articles according to their subject areas and their percentages in the total number of articles. The classification of the articles depending on the subject areas was based on the systematic taxonomy provided

TABLE 3. Subject areas of the articles.

Subject Area	N	%
Mathematics	62,394	87.28
Biochemistry, Genetics, and Molecular Biology	62,015	86.75
Computer Science	42,532	59.49
Agricultural and Biological Sciences	28,985	40.54
Immunology and Microbiology	18,438	25.79
Environmental Science	7,139	9.99
Neuroscience	7,139	9.99
Medicine	4,897	6.85
Engineering	4,251	5.95
Chemistry	3,076	4.30
Pharmacology, Toxicology, and Pharmaceutics	773	1.08
Social Sciences	697	0.97
Health Professions	401	0.56

by Scopus, which was designed by considering the scientific background of the articles and related journals. As the articles dealt with several areas, the sum of percentages exceeded 100%. The table also reveals the interdisciplinary nature of bioinformatics. This table shows that the majority of the contributions came from mathematics (87.28%), biochemistry, genetics and molecular biology (86.75%), and computer science (59.49%).

B. TOPIC MODELING ANALYSIS

In order to discover the emerging topics of bioinformatics studies (RQ2), a topic modeling analysis was conducted. After applying the LDA topic modeling procedure to the created corpus, 24 main topics were found, as listed in Table 4. The top keywords classified for each topic by the LDA model are also given in this table. The topic names were given by considering all the classified keywords under each topic; however, as the ratio was higher for the first keywords, the first keywords were primarily considered when naming each topic. The rate of the topics indicates the frequency of occurrence of each topic. Accordingly, when all articles published between 1971 and 2020 are considered, the topic ''Statistical Estimation'' has the highest ratio, indicating that the occurrence of this topic is the highest in the articles.

C. TEMPORAL TOPIC TRENDS

In order to better understand how the trends and evolutions in each topic change over time (RQ3), details of the discovered topics for each five-year period starting from 1971 until 2020 were calculated and presented. In this context, the acceleration values of each topic were calculated by taking into account the annual changes in the rates of the topics. Their averages for each five-year period between 1971 and 2020 are given in Fig. 2, where the dashed (blue) lines show the acceleration values calculated for each period and the straight arrow (red) lines show their linear trend-line, providing some additional information for their near-future trends. Fig. 2 shows that whereas trends of some topics like ''Protein Structure'', ''Population Evolution'', ''Equations'', and ''Metabolic Reactions'' are declining, those of ''Prediction'', ''Gene Expression'', and ''Genomic Data'' are accelerating positively.

Analysis of the average acceleration values of the topics from 1971 to 2020 shows that the topics of ''Data Analysis Tools'', ''Prediction'', ''Genomic Data'', and ''Gene Expression'' occupy the first four ranks with the highest positive acceleration values of 0.15, 0.14, 0.13, and 0.12, respectively (Fig. 3).

On the other hand, the topics ''Equations'' and ''Membrane Transport'', with the lowest acceleration values (negatively accelerated) of −0.41 and −0.34, respectively (Fig. 3). These results indicated the emergence of new topics through prediction-based analyses and tools on biological data.

D. DEVELOPMENTAL STAGES OF THE FIELD

In order to better understand the developmental stage of each topic between 1971 and 2020 (RQ4), the top five topics of each period were also analyzed (Fig. 4). In this figure, the topics in bold are those not included in the list of the previous terms, and the topics are ordered according to their volume from the highest to the lowest. As seen from Fig. 4, ''Tumor Cells'' and ''Metabolic Reactions'' were the most frequently studied topics until 1976; however, starting from 1971 until 2001, topics like ''Equations'', ''Membrane Transport'', ''Species Population'', and ''Population Evolution'' became more dominant. Furthermore, after 2001, a variety of topics began to appear in the top-5 list, such as ''Statistical Estimation'', ''Data Analysis Tools'', and ''Algorithms'', indicating the tendency of the studies to perform data analysis on biological research. After 2016, the topic ''Prediction'' became one of the dominant topics. Hence, the topics in bioinformatics literature had evolved from population-based understandings into prediction-based classifications, machine learning algorithms, and artificial intelligence methods for improved predictions on biological data. By considering the top-five topics in each period, the developmental stages of the field were divided into seven stages. These stages were named by the authors as Newborn (1970–1975), Infancy (1975–1985), Childhood (1985–2000), Adolescence (2000–2005), Adulthood (2005–2010), Mature (2010–2015), and Wisdom (2015–2020).

E. RECENT TOPIC TRENDS

In order to better understand the recent trends (RQ5), the acceleration values of all topics during the last 5-year period (2016–2020) were also analyzed. As seen in Fig. 5, although some of the topics had lost their popularity, the topic ''Prediction'' had the highest recent acceleration value (0.52). On the other hand, several topics showed negative accelerations. This result indicated that prediction-related studies would dominate future bioinformatics research.

V. DISCUSSION

In this study, all major articles published from 1971 until 2020 in the field of bioinformatics were analyzed to better understand the main topics, developmental stages, trends,

TABLE 4. Discovered topics, their keywords and rates.

Topic Name	Keywords	Rate
Statistical Estimation	data*; estim*; statist*; analysi*; measur*; sampl*; test*; paramet*; set*; simul*; distribut*; perform*; variabl*; experi*; effect*	7.59
Equations	system*; equat*; dynam*; time*; paramet*; stochast*; rate*; distribut*; process*; solut*; popul*; function*; stabil*; period*; condit*	7.50
Data Analysis Tools	data*; tool*; analysi*; softwar*; visual*; databas*; user*; implement*; develop*; web*; packag*; univers*; integr*; comput*; allow*	6.87
Algorithms	algorithm*; tree*; comput*; cluster*; optim*; effici*; time*; phylogenet*; set*; distanc*; graph*; search*; sequenc*; program*; space*	5.17
Protein Sequence	protein*; sequenc*; structur*; align*; predict*; function*; similar*; secondari*; rna*; domain*; databas*; score*; famili*; acid*; peptid*	4.99
Gene Expression	gene*; express*; data*; cancer*; analysi*; identifi*; microarray*; cluster*; differenti*; pathway*; set*; profil*; function*; biolog*; dataset*	4.70
Ontology	inform*; ontolog*; annot*; data*; system*; biolog*; biomed*; knowledg*; extract*; databas*; semant*; text*; term*; develop*; mine*	4.56
Prediction	predict*; featur*; learn*; classif*; perform*; machin*; classifi*; accuraci*; select*; data*; train*; dataset*; base*; algorithm*; improv*	4.41
Protein Interaction	network*; interact*; protein*; function*; biolog*; complex*; gene*; regulatori*; infer*; data*; protein-protein*; modul*; structur*; topolog*; graph*	4.38
Genomic Data	sequenc*; genom*; read*; variant*; assembl*; data*; genet*; map*; associ*; genotyp*; detect*; snps*; tool*; haplotyp*; metagenom*	4.32
Protein Structure	protein*; structur*; bind*; interact*; site*; conform*; residu*; energi*; molecular*; acid*; complex*; function*; peptid*; molecul*; predict*	4.11
Population Evolution	popul*; select*; evolut*; mutat*; evolutionari*; genet*; individu*; cooper*; fit*; strategi*; game*; theori*; rate*; effect*; adapt*	4.07
Membrane Transport	membran*; transport*; flow*; mechan*; channel*; muscl*; cell*; effect*; potenti*; forc*; simul*; concentr*; ion*; increas*; tissu*	4.04
Genome Sequence	genom*; sequenc*; gene*; dna*; region*; speci*; code*; chromosom*; repeat*; evolut*; methyl*; base*; codon*; compar*; organ*	3.93
Species Population	speci*; popul*; competit*; effect*; plant*; dynam*; growth*; predat*; dispers*; communiti*; ecolog*; spatial*; resourc*; increas*; environ*	3.73
Signaling Regulatory	signal*; regul*; system*; activ*; pathway*; respons*; dynam*; cell*; mechan*; control*; protein*; feedback*; simul*; biolog*; level*	3.38
Tumor Cells	cell*; tumor*; growth*; cancer*; develop*; tissu*; respons*; differenti*; immun*; cellular*; mathemat*; stem*; prolifer*; mechan*; experiment*	3.28
Infectious Diseases	infect*; diseas*; virus*; host*; epidem*; transmiss*; vaccin*; viral*; pathogen*; control*; resist*; dynam*; immun*; popul*; effect*	3.01
Transcription	transcript*; site*; motif*; mirna*; bind*; gene*; sequenc*; factor*; regul*; rna*; regulatori*; target*; predict*; splice*; identifi*	3.00
Neuron Activity	neuron*; activ*; brain*; neural*; respons*; dynam*; spike*; input*; synapt*; inform*; behavior*; network*; mechan*; connect*; learn*	2.86
Metabolic Reactions	metabol*; reaction*; enzym*; pathway*; flux*; metabolit*; analysi*; product*; substrat*; concentr*; kinet*; bacteri*; growth*; experiment*; system*	2.70
Drug Compounds	drug*; target*; compound*; chemic*; molecular*; screen*; activ*; inhibitor*; predict*; molecul*; design*; structur*; develop*; effect*; potenti*	2.56
Clinical Diseases	diseas*; patient*; clinic*; associ*; treatment*; medicin*; risk*; health*; medic*; effect*; human*; data*; diagnosi*; phenotyp*; disord*	2.45
Spatial Patterns	imag*; pattern*; segment*; spatial*; analysi*; structur*; shape*; detect*; develop*; map*; local*; process*; base*; microscopi*; format*	2.37

and future directions of bioinformatics literature. This study offers several findings that provide very valuable information for researchers engaged in the field and for decision-makers in related areas. These findings are summarized under the headings below.

A. INTERDISCIPLINARY NATURE OF THE FIELD

By analyzing earlier studies, this current study has highlighted the interdisciplinary nature of the field, which was also emphasized by earlier studies. However, as additional information, this paper has also listed the level of contribution from other disciplines over the last 50 years. The results have indicated that there is a heavy contribution from fields like mathematics, biochemistry, genetics and molecular sciences, informatics, and computer science (Table 3). This situation is also reflected in the discovered topics. For example, the topic ''Equations'' can be considered a contribution from mathematics, ''Statistical Estimation'' from statistics, and ''Data Analysis Tools'', ''Prediction'', and ''Algorithms'' from computer science. Similarly, the topic ''Drug Compounds'' can be considered a contribution from biochemistry and ''Genomic Data'' and ''Genome Sequence'' from genetics.

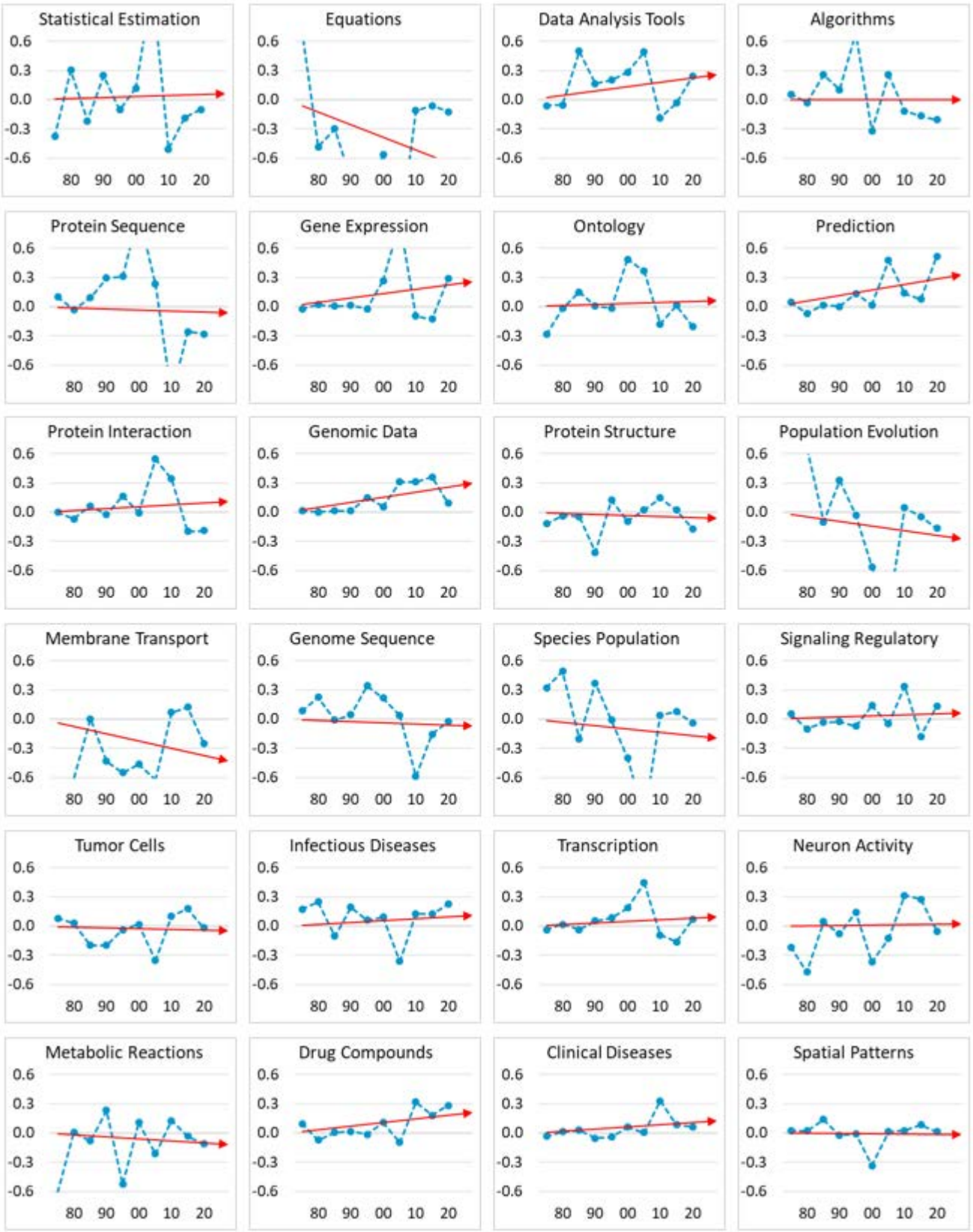

FIGURE 2. Acceleration graphs of the discovered topics.

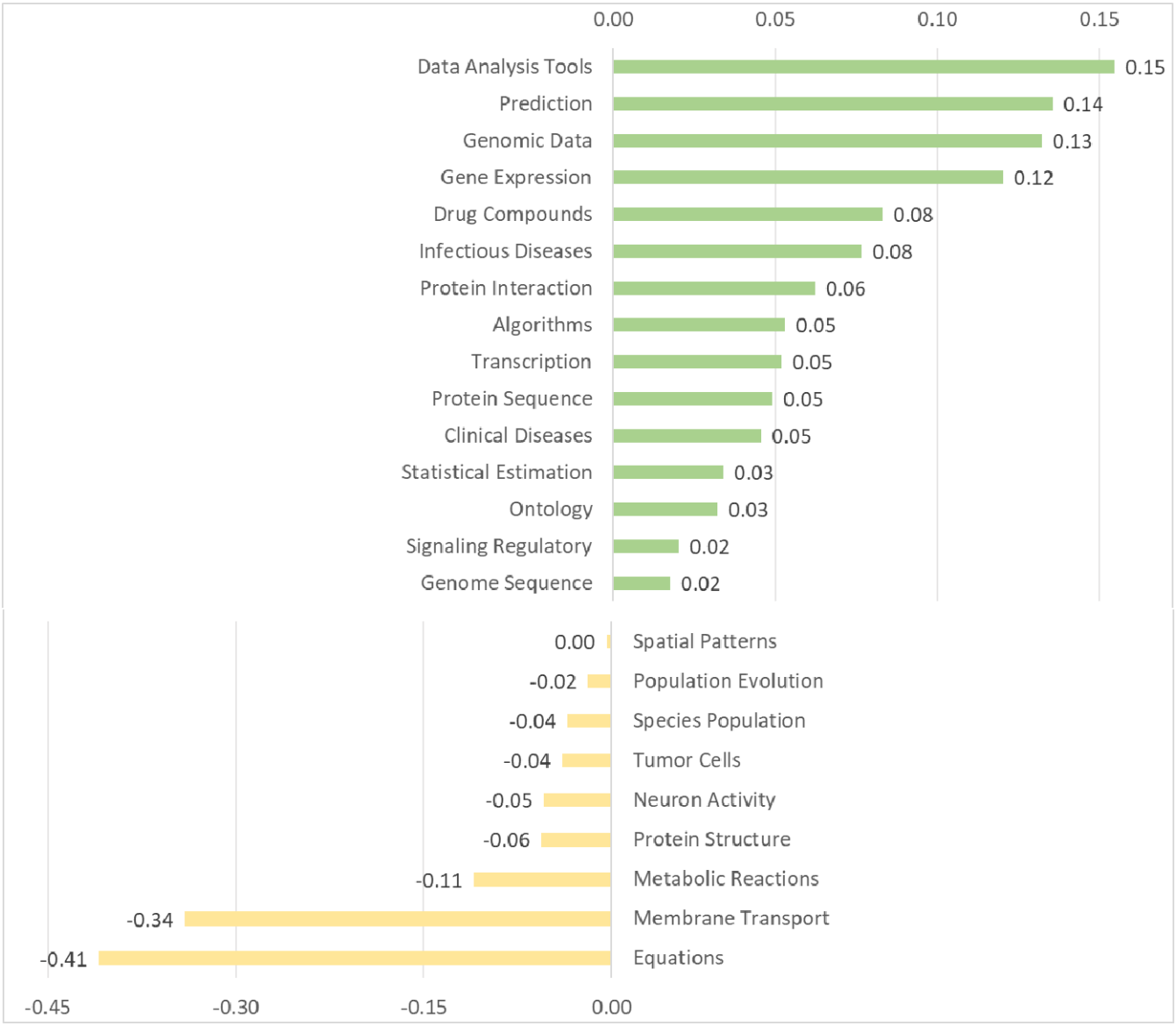

FIGURE 3. Average acceleration of each topic from 1971 to 2020.

B. TEMPORAL LANDSCAPE OF THE TOPICS AND TRENDS

According to the results, 24 topics were identified in the 71,490 articles through text mining analysis. The topics listed have similarities with earlier studies. For example, the topics discovered in this study, namely ''Gene Expression'', ''Ontology'', and ''Machine Learning'' (Prediction) were also reported in [23]. Their results indicated a pattern for these topics similar to our results. Both studies indicated an increased number of studies on the topic ''Gene Expression'' between 1998 and 2015. However, because this current study also covered the later years until 2020, as additional information, it showed a decrease in this topic after 2015. Similar results were also reported by both studies for the topic ''Ontology''. In both studies, there were an increased number of studies on this topic from 1998 until 2010, whereas a sharp reduction was seen in 2010. Both studies also reported a similar increase for the topic ''Prediction'' (''machine learning'' in [23]). However, in terms of methods, the other study [23] was performed on an experimental dataset created using articles containing only the term ''bioinformatics''. In addition, in that study, it was determined that the preprocessing steps were limited and only the top ten topics were obtained by superficially applying the topic modeling. The top 25 keywords revealed in their study (which we interpreted as topics) were chosen using a manual approach based on frequency of word occurrence.

The topics ''Algorithms'', ''Metabolic Reactions'' (Pathways), ''Gene Expression'', ''Protein Structure'', ''Protein Interaction'', and ''Drug Compounds'' (Chemical Drugs) are the same or very similar to those listed in [14]. Moreover,

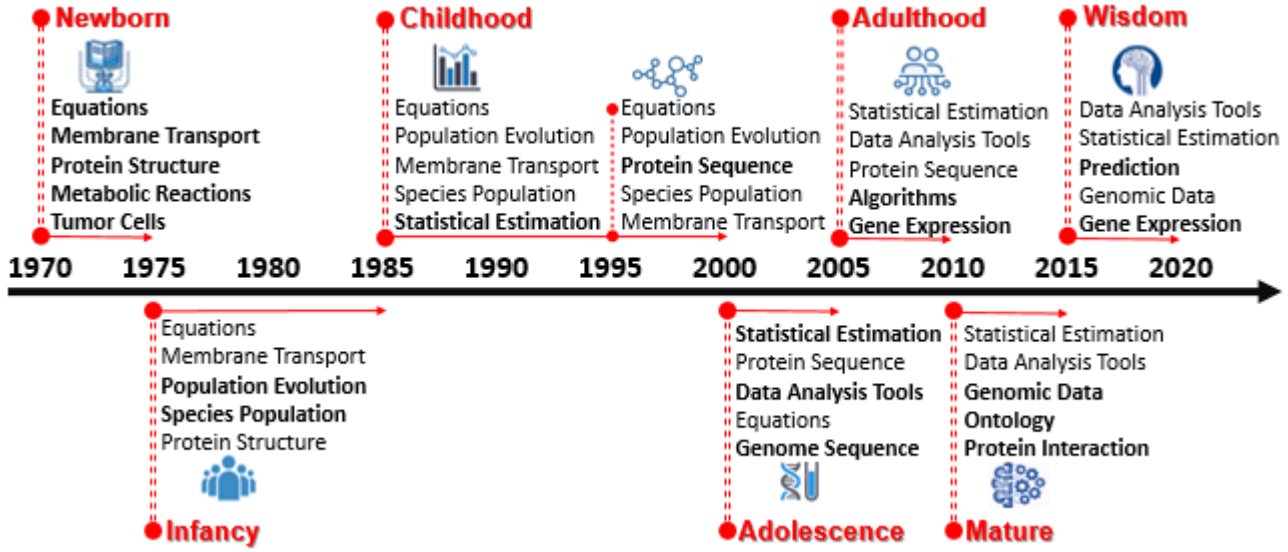

FIGURE 4. The seven developmental stages of bioinformatics studies.

they reported a decline in the topic ''Algorithms'', which was also detected in our results (Fig. 2). They also reported a peak in the number of studies on the topic ''Protein Structure'' around 1996-97 and a decrease until 1999 [14]. A similar pattern was also detected for this topic in our results (Fig. 2). Similarly, in their study, the topic ''Gene Expression'' peaked after 2000 and decreased after 2006 [14], a trend also found in our results (Fig. 2). They reported an increase in the number of studies on topics related to data mining [14], which can be concluded as the reason for the very high acceleration value detected for the topic ''Prediction'' in our current study. Because the current study was conducted almost 10 years after [14], these similarities can be considered as indicators that validate the results of both studies. On the other hand, there are important technical differences in their methodology compared to ours. Namely, the experimental dataset used in their study included only articles from the Journal of Bioinformatics, and therefore, the analysis carried out had a very limited perspective. In addition, in that study, data preprocessing steps were not addressed, the dominance of the topics obtained with LDA was not calculated, and the presented topics were selected manually [14]. The findings similar to those of earlier studies are indicators for validating the results of this current study and those of the earlier studies. However, through an improved methodology, our current study provides a wider perspective for the field of bioinformatics. Moreover, the proposed methodology for a systematic and objective creation of a corpus for bioinformatics literature is important because it enables regular monitoring of the field using a similar approach.

C. SEVEN DEVELOPMENTAL STAGES FROM BIRTH TO WISDOM

The developmental stages of the field can be analyzed from many perspectives. As the bioinformatics field has an interdisciplinary nature, several other fields have contributed to it from different perspectives, which has impacted the topics of the field. By considering the developmental stages of bioinformatics literature, [22] named the period between 2002 and 2006 as ''Adolescence'' and between 2007 and 2011 as ''Adulthood''. As seen from Fig. 4, our results also indicated a change during these periods that is consistent with [22], and we used their naming strategy in our study as well. Accordingly, during these stages, the topics ''Protein Sequence'', ''Data Analysis Tools'', and ''Statistical Estimation'' were common in the top-five topics list, and ''Equations'' and ''Genome Sequence'' were replaced by ''Algorithms'' and ''Gene Expression'' after 2006. In addition, we found another change after 2006. The topics ''Genomic Data'', ''Ontology'', and ''Protein Interaction'' were then recorded in the top-five topics list along with the topics ''Data Analysis Tools'' and ''Statistical Estimation'' until 2016; however, after 2016, a shift towards ''Prediction'' and ''Gene Expression'' was detected. Another important finding was that during the 30 years starting from 1971 until 2001, only nine topics were included in the top-five topics list, whereas during the last 20 years between 2001 and 2020, 14 topics found a place in the list. This result indicated the rapid pace at which the field was evolving and changing. Additionally, by expanding [22], the evolution of the bioinformatics field can be divided into seven stages: Newborn (1971-1975), Infancy (1976-2000), Adolescence (2001-2005), Adulthood (2006-2010), Mature (2011-2016), and Wisdom (2016-2020).

D. PROGRESSION OF THE TWO MAIN TRAJECTORIES

As its name indicates, bioinformatics research has been mainly influenced by the fields of biology and informatics. In our current study, we demonstrated the contributions of

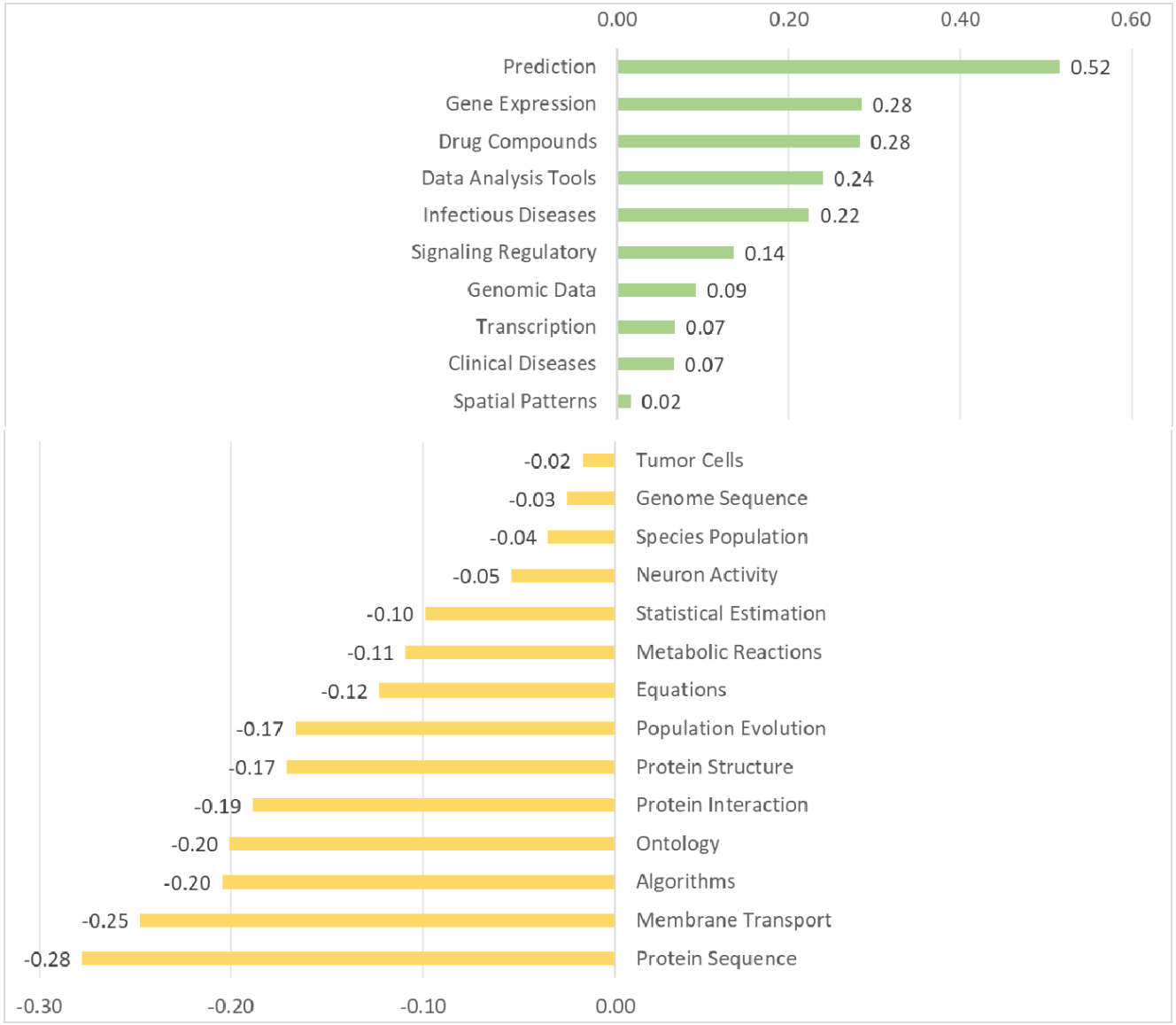

FIGURE 5. Acceleration percentages of topics from 2016 to 2020.

each field during the developmental stages of bioinformatics. Fig. 6 shows the top-five topics of each period classified according to these two trajectories. In this figure, the topics that emerged for the first time in the given developmental stage are given in bold. From Fig. 6, the influence of the biology and informatics trajectories of bioinformatics can be explicitly seen in each developmental stage. When the informatics trajectory is considered, using computational methods and ''Data Analysis Tools'', the studied topics evolved from simpler concepts like ''Equations'' and ''Statistical Estimation'' towards more complex concepts like ''Prediction''. A similar development can be seen in the biology trajectory, as well. The studied topics in the early ages were generally related to some basic concepts like ''Tumor Cells'' and ''Metabolic Reaction''; however, over time, they evolved towards more complex concepts like ''Genomic Data'' and ''Gene Expression''.

The collaboration of these two disciplines seems to have closely affected the developmental stages of the field from the ''Newborn'' to the ''Wisdom'' stage. Hence, during its early phases, including the Newborn and Infancy stages, the field was mainly dominated by biological studies. After the Childhood stage, informatics studies, by using tools and computational methods for the analysis of biological data, began to provide higher contributions, through topics like ''Algorithms'', ''Data Analysis Tools'', ''Ontology'', and even ''Prediction''.

E. FUTURE OUTLOOKS FOR BIOINFORMATICS: WHERE TO GO FROM HERE?

Considering the results, the rising trend in the topic of ''Prediction'' emphasizes that machine learning-based advanced techniques provide an effective methodology for analyzing complex biological data. As also highlighted by earlier

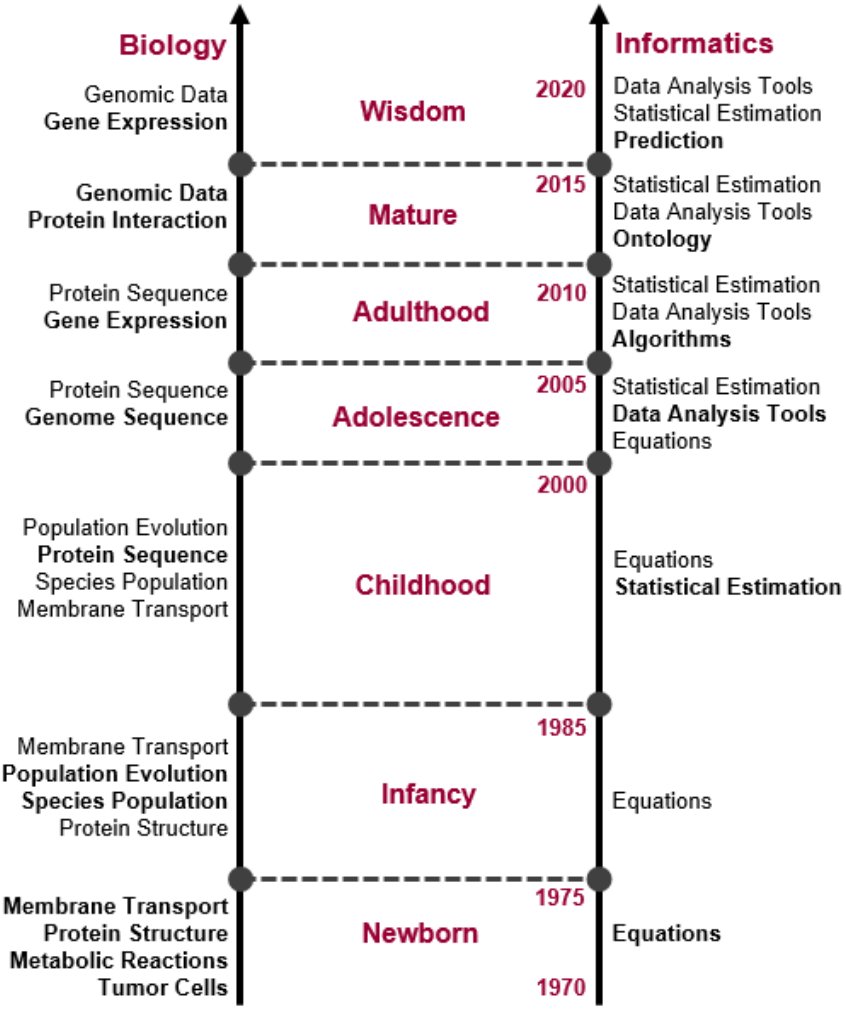

FIGURE 6. Biology and informatics trajectories.

studies, the amount of biomedical data is significant, and machine learning is a widely-used method for analyzing this data [10], [23], [46]. The importance of prediction-based analysis in bioinformatics is also emphasized by another study showing the developmental trends of miRNA tools in bioinformatics [24]. Consistent with these earlier studies, our results indicated the future direction of bioinformatics via the futurecasting method and prediction-based analysis. Fig. 6 shows that as the field of bioinformatics progresses towards the ''Mature'' stage, the expectations from the informatics concepts are also more likely to get clearer, and accordingly, the number of topics from this trajectory will increase. We can conclude from this figure that in the future, the informatics trajectory will contribute more to the field of bioinformatics.

VI. LIMITATIONS OF THE STUDY; FUTURE WORK

As with all research, this study has some limitations. In this study, the corpus was developed using 20 journals and conferences based on their H-index values. This approach provided an opportunity for creating an appropriate corpus for the study; however, on the other hand, more recent journals and organizations may not have been included in the corpus. Hence, when creating their corpora, future studies should develop approaches for including these newly established data sources. Additionally, although several phenomena have influenced the field of bioinformatics, such as the research on BLAST and the human genome and the advent of NGS, no articles on these topics were included in this current study. In future studies, this methodology could be improved by including the contributions of these accomplishments in order to form a deeper understanding of the field.

Alternatively, instead of using the major publication sources (e.g., major journals, conferences, and events in the bioinformatics field), the proposed methodology could create the corpus by using some specific keywords of the field. For example, in order to reveal trends for narrower, specific subfields such as ''light microscopic imaging'', ''computational

mass spectrometry'', ''cell segmentation'', ''oxidation'', and ''3D nuclear architecture'' or for very new emerging trends such as ''AlphaFold'' by DeepMind, the corpus could be created through related keyword searches. In addition, a website could be created in order to provide services for corpus development and to offer an analysis of the corpus data. Such an application would be very helpful when used to regularly monitor and compare the general topics and their trends. Moreover, some specific topics of the field and their trends could also be analyzed through such an application. However, for now, the corpus of this study and other details and specific sub-fields have been shared for further analysis as supplementary file attachments to this study.

Furthermore, new supportive approaches could be developed using different data processing methods, preprocessing steps, and topic modeling algorithms to adapt new topic models to different textual datasets. Different holistic analyses could be carried out by using these approaches together. The application of different topic modeling algorithms such as Hierarchical Latent Dirichlet Allocation (HLDA), Non-Negative Matrix Factorization (NMF), Dirichlet Multinomial Regression (DMR), Hierarchical Dirichlet Process (HDP), Correlated Topic Model (CTM), Dynamic Topic Model (DTM), and Gaussian Latent Dirichlet Allocation (GLDA) to various textual corpora, and a comparative analysis of the results obtained with these algorithms would expand this methodology and provide important contributions to future studies. Many potential avenues are envisaged for future work in this constantly evolving field.

VII. CONCLUSION

To begin with, this study provided an objective and systematic approach to the creation of a corpus of bioinformatics studies. The trends emerging in the field were analyzed by applying an approach to the tools similar to that developed in an earlier study [26]. Such tools are important to support the researchers engaged in the field with continuous feedback from the past. Secondly, when the bioinformatics studies were analyzed, two main trajectories could be recognized: biology and informatics. As was also highlighted by [1], the bioinformatics tools indicated under the informatics trajectory defined in the current study included all the methods, approaches, and developments related to the comparative analysis of biological data. On the other hand, the biological studies dealt with all the methods and approaches to collect, identify, and distinguish the biological data to be analyzed for further improvements and developments in the field. These trajectories can be validated through the disciplines involved in the bioinformatics literature, as listed in Table 3. From this table, it can be concluded that disciplines like ''Mathematics'', ''Computer Science'', and ''Engineering'' provide contributions to developments in the informatics track of the field. On the other hand, disciplines like ''Biochemistry'', ''Genetics and Molecular Reactions'', ''Agricultural and Biological Sciences'', ''Immunology and Microbiology'', ''Environmental Science'', ''Neuroscience'', ''Medicine'', ''Chemistry'', ''Pharmacology, Toxicology, and Pharmaceutics'', ''Social Sciences'', and ''Health Professions'' contribute to the biological track of the field. Similar developmental tracks can also be seen in Fig. 6. Accordingly, the results indicate that the development of bioinformatics studies has been in a close relationship with technological advances; therefore, future studies to be carried out in the field can be expected to focus more heavily on the predictive analysis of biological knowledge and data.

According to the results, it can be concluded that, after 2000, the improvements in the topics ''Statistical Estimation'' and ''Data Analysis Tools'' laid the foundation for the developments in ''Prediction''-based studies in the field (Fig. 4). Consistent with these results, the acceleration value for the topic ''Prediction'' has been very high for the last five years (Fig. 5), which indicates that future bioinformatics studies will be closely related to prediction-based analysis. Thus, it can be concluded that bioinformatics is a pioneering field that lays the foundation for the application of innovative computational methods to biological data. This study has attempted to provide a more holistic picture of the field by considering major studies conducted from 1971 to 2020. The authors believe that the contributions provided by this picture will serve as a guide for researchers and decision-makers in the field of bioinformatics.

REFERENCES

[1] I. Y. Abdurakhmonov, *Bioinformatics: Basics, Development, and Future*. Rijeka, Croatia: InTech, 2016.
[2] N. M. Luscombe, D. Greenbaum, and M. Gerstein, ''What is bioinformatics? A proposed definition and overview of the field,'' *Methods Inf. Med.*, vol. 40, no. 4, pp. 346–358, 2001.
[3] P. Hogeweg, ''The roots of bioinformatics in theoretical biology,'' *PLoS Comput. Biol.*, vol. 7, no. 3, 2011, Art. no. e1002021.
[4] A. Lesk, *Introduction to Bioinformatics*. Oxford, U.K.: Oxford Univ. Press, 2019.
[5] A. D. Baxevanis, G. D. Bader, and D. S. Wishart, *Bioinformatics*. Hoboken, NJ, USA: Wiley, 2020.
[6] X. Zhang, X. Zhou, and X. Wang, ''Basics for bioinformatics,'' in *Basics of Bioinformatics*. Berlin, Germany: Springer, 2013, pp. 1–25.
[7] A. Youssef and A. Rich, ''Exploring trends and themes in bioinformatics literature using topic modeling and temporal analysis,'' in *Proc. IEEE Long Island Syst., Appl. Technol. Conf. (LISAT)*, May 2018, pp. 1–6.
[8] B. Hesper and P. Hogeweg, ''Bioinformatica: Een werkconcept,'' *Kameleon*, vol. 1, no. 6, pp. 28–29, 1970.
[9] J. D. Wren, ''Bioinformatics programs are 31-fold over-represented among the highest impact scientific papers of the past two decades,'' *Bioinformatics*, vol. 32, no. 17, pp. 2686–2691, Sep. 2016.
[10] X. Jiang, J. Zhao, W. Qian, W. Song, and G. N. Lin, ''A generative adversarial network model for disease gene prediction with RNA-seq data,'' *IEEE Access*, vol. 8, pp. 37352–37360, 2020.
[11] M. Ishaq, A. Khan, M. Khan, and M. Imran, ''Current trends and ongoing progress in the computational alignment of biological sequences,'' *IEEE Access*, vol. 7, pp. 68380–68391, 2019.
[12] A. Kastrin and D. Hristovski, ''Scientometric analysis and knowledge mapping of literature-based discovery (1986–2020),'' *Scientometrics*, vol. 126, pp. 1415–1451, Feb. 2021.
[13] M. Song and S. Y. Kim, ''Detecting the knowledge structure of bioinformatics by mining full-text collections,'' *Scientometrics*, vol. 96, no. 1, pp. 183–201, Jul. 2013.
[14] Y. Wu, H. Wang, M. Feng, and J. Pei, ''Evolving research trends in bioinformatics,'' in *Proc. 4th Int. Conf. Bioinf. Biomed. Eng.*, 2010, pp. 1–4.

[15] J. P. S. Manoharan, A. Kanagavel, B. Muthuchidambaram, and A. Kumaravel, "Bioinformatics research—An informetric view," in *Proc. Int. Conf. Inf. Commun. Manage.*, vol. 16, 2011, pp. 199–204.

[16] G. E. Heo, K. Y. Kang, M. Song, and J.-H. Lee, "Analyzing the field of bioinformatics with the multi-faceted topic modeling technique," *BMC Bioinf.*, vol. 18, no. S7, p. 251, May 2017.

[17] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, "Big data application in biomedical research and health care: A literature review," *Biomed. Informat.*, vol. 8, Jan. 2016, Art. no. S31559.

[18] A. J. van Altena, P. D. Moerland, A. H. Zwinderman, and S. D. Olabarriaga, "Understanding big data themes from scientific biomedical literature through topic modeling," *J. Big Data*, vol. 3, no. 1, p. 23, Dec. 2016.

[19] D. S. M. da Silva, W. M. C. da Silva, G. RuiZhe, A. P. Bernardi, A. M. Mariano, and M. Holanda, "Big data trends in bioinformatics," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 1862–1867.

[20] M. Song, G. E. Heo, and S. Y. Kim, "Analyzing topic evolution in bioinformatics: Investigation of dynamics of the field with conference data in DBLP," *Scientometrics*, vol. 101, no. 1, pp. 397–428, Oct. 2014.

[21] S. K. Patra and S. Mishra, "Bibliometric study of bioinformatics literature," *Scientometrics*, vol. 67, no. 3, pp. 477–489, Jun. 2006.

[22] C. A. Ouzounis, "Rise and demise of bioinformatics? Promise and progress," *PLoS Comput. Biol.*, vol. 8, no. 4, Apr. 2012, Art. no. e1002487.

[23] P. Hahn, A. Mohanty, and S. D. Manda, "What's hot and what's not?-exploring trends in bioinformatics literature using topic modeling and keyword analysis," in *Proc. Int. Symp. Bioinf. Res. Appl.* Cham, Switzerland: Springer, 2017, pp. 279–290.

[24] J. Guan and X. Gao, "Comparison and evaluation of Chinese research performance in the field of bioinformatics," *Scientometrics*, vol. 75, no. 2, pp. 357–379, May 2008.

[25] C. Perez-Iratxeta, M. A. Andrade-Navarro, and J. D. Wren, "Evolving research trends in bioinformatics," *Briefings Bioinf.*, vol. 8, no. 2, pp. 88–95, Dec. 2006.

[26] S. Kavvadias, G. Drosatos, and E. Kaldoudi, "Supporting topic modeling and trends analysis in biomedical literature," *J. Biomed. Informat.*, vol. 110, Oct. 2020, Art. no. 103574.

[27] F. Gurcan and N. E. Cagiltay, "Research trends on distance learning: A text mining-based literature review from 2008 to 2018," *Interactive Learn. Environ.*, vol. 7, pp. 1–22, Sep. 2020.

[28] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012.

[29] F. Gurcan, N. E. Cagiltay, and K. Cagiltay, "Mapping human–computer interaction research themes and trends from its existence to today: A topic modeling-based review of past 60 years," *Int. J. Hum.-Comput. Interact.*, vol. 37, no. 3, pp. 267–280, Feb. 2021.

[30] Y. Lu, Q. Mei, and C. Zhai, "Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA," *Inf. Retr.*, vol. 14, no. 2, pp. 178–203, 2011.

[31] P. Mongeon and A. Paul-Hus, "The journal coverage of web of science and scopus: A comparative analysis," *Scientometrics*, vol. 106, no. 1, pp. 213–228, Jan. 2016.

[32] F. Gurcan, O. Ozyurt, and N. E. Cagiltay, "Investigation of emerging trends in the e-learning field using latent Dirichlet allocation," *Int. Rev. Res. Open Distrib. Learn.*, vol. 22, no. 2, pp. 1–18, 2021.

[33] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Inf. Process. Manag.*, vol. 50, pp. 104–112, Jan. 2014.

[34] M. Konrad. (2017). *Text Mining and Topic Modeling Toolkit*. Accessed: Jan. 21, 2022. [Online]. Available: https://pypi.org/project/tmtoolkit/

[35] *GitHub—WZB Social Science Center/Tmtoolkit: Text Mining and Topic Modeling Toolkit for Python With Parallel Processing Power*. Accessed: Jan. 13, 2022. [Online]. Available: https://github.com/WZBSocialScienceCenter/tmtoolkit

[36] *Tmtoolkit: Text Mining and Topic Modeling Toolkit-Tmtoolkit Documentation*. Accessed: Jan. 16, 2022. [Online]. Available: https://tmtoolkit.readthedocs.io/en/latest/index.html

[37] M. Porter, *Snowball: A Language for Stemming Algorithms*. Chennai, India: Snowball, 2001.

[38] D. M. Blei, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, nos. 4–5, pp. 993–1022, 2003.

[39] F. Gurcan and N. E. Cagiltay, "Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling," *IEEE Access*, vol. 7, pp. 82541–82552, 2019.

[40] D. M. Blei and J. D. Lafferty, "Correction: A correlated topic model of science," *Ann. Appl. Statist.*, vol. 1, no. 2, p. 634, Dec. 2007.

[41] *Tomotopy Topic Modeling Package*. Accessed: Feb. 27, 2022. [Online]. Available: https://bab2min.github.io/tomotopy/v0.12.2/en/

[42] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Inf. Syst.*, vol. 94, Dec. 2020, Art. no. 101582.

[43] F. Gurcan, "Extraction of core competencies for big data: Implications for competency-based engineering education," *Int. J. Eng. Educ.*, vol. 35, no. 4, pp. 1110–1115, 2019.

[44] C. Li, S. Feng, Q. Zeng, W. Ni, H. Zhao, and H. Duan, "Mining dynamics of research topics based on the combined LDA and WordNet," *IEEE Access*, vol. 7, pp. 6386–6399, 2019.

[45] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 262–272.

[46] Z. Tao, B. Dong, Z. Teng, and Y. Zhao, "The classification of enzymes by deep learning," *IEEE Access*, vol. 8, pp. 89802–89811, 2020.



FATIH GURCAN received the B.S. degree in statistics and computer science and the M.S. and Ph.D. degrees in computer engineering from Karadeniz Technical University, Trabzon, Turkey, in 2001, 2009, and 2017, respectively. He worked as an Instructor with the Department of Informatics, Karadeniz Technical University, from 2001 to 2014, where he has been with the Center for Research and Application in Distance Education, as an Instructor, since 2015. His research interests include trend analysis, sentiment analysis, statistical topic modeling, engineering education, data mining, machine learning, big data analytics, and text mining.



NERGIZ ERCIL CAGILTAY received the degree in computer engineering and the Ph.D. degree in instructional technologies from Middle East Technical University, Turkey. She worked with Commercial and Government Organizations as a Project Manager for more than eight years, in Turkey. She also worked with the Indiana University Digital Library Program as a System Analysis and a Programmer for four years. Since 2003, she has been with the Software Engineering Department, Atilim University, Turkey, as an Associate Professor. Her research interests include information systems, medical information systems, engineering education, instructional systems technologies, distance education, e-learning, and medical education.