Big Data Research 27 (2022) 100289

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

Big Data in Forecasting Research: A Literature Review

Ling Tang [a,b], Jieyi Li [a], Hongchuan Du [c], Ling Li [d,*], Jun Wu [a], Shouyang Wang [e]

[a] School of Economics and Management, Beijing University of Chemical Technology, Beijing 100029, China
[b] School of Economics and Management, Beihang University, Beijing 100191, China
[c] School of Science, Beijing University of Chemical Technology, Beijing 100029, China
[d] International School of Economics and Management, Capital University of Economics and Business, Beijing 100070, China
[e] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 25 April 2021
Received in revised form 17 October 2021
Accepted 30 October 2021
Available online 9 November 2021

Keywords:
Big data
Forecasting
Literature review
Prediction models
Information

ABSTRACT

With the boom in Internet techniques and computer science, a variety of big data have been introduced into forecasting research, bringing new knowledge and improving prediction models. This paper is the first attempt to conduct a literature review on full-scale big data in forecasting research. By source, big data in forecasting research fell into user-generated content data (from the users on social media in texts, photos, etc.), device-monitored data (by meteorological monitors, smart meters, GPS, etc.) and activity log data (for web searching/visiting, online/offline marketing, clinical treatments, laboratory experiments, etc.). Different data types, bearing distinctive information and characteristics, dominated different forecasting tasks, required different analysis technologies and improved different forecasting models. This survey provides an overall review of big data-based forecasting research, details what (regarding data types and sources), where (forecasting hotspots) and how (analysis and forecasting methods used) big data improved prediction, and offers insights into future prospects.

1. Introduction

With the boom in Internet techniques and computer science, a variety of data in unstructured or semi-unstructured formats have emerged and accumulated, forming *big data* and describing the world from different perspectives [1]. Even without a uniform definition yet, big data have generally been considered to be characterized by the 5V, i.e., *Volume*, *Variety*, *Velocity*, *Value* and *Veracity* [2,3]. Over the past decades, big data have been introduced into diverse research fields, bringing encouraging innovations to the associated theories and technologies [4,5]. On the one hand, informative big data have provided new information and knowledge, supporting a new or better understanding for the targeted issues and thereby challenging and even reshaping the basic theories that were based on traditional data [6]. On the other hand, in a different (unstructured or semi-unstructured) format, these big data have called for a substantial renovation of the processing and analysis techniques and even given birth to new methodologies, particularly in the field of computer science [7].

With the help of big data, forecasting research, a fundamental scientific research covering nearly ever research domain and capturing both history and future dynamics, has experienced large innovation in recent years [8,9]. Generally, the big data that have been applied to forecasting research came from three main sources: Internet users, monitoring devices and activity logs [10–12]. First, the Internet has provided a broad platform, in terms of social media, allowing its users to share individual information in the formats of online texts, online photos, etc., forming user-generated content (UGC) data [13] and bringing new, predictive knowledge (e.g., public opinions, emotions and attention) as an important input to forecasting models. Second, with the rapid development of the Internet of things (IoT), helpful devices (e.g., meteorological monitors, smart meters, global position systems (GPS), and other sensors or monitors) have been employed to directly track specific targets in a monitor-level real-time way, forming device-monitored data (hereafter abbreviated as device data) and enhancing the spatiotemporal resolution of prediction [14]. Third, due to the updates of data processing and storage technologies, detailed activities or operations (such as web searching and visiting, online and offline marketing, clinical treatments and lab experiments) can be recorded in terms of activity log data (regarding who, when, where and what for each activity), facilitating an intensive and extensive exploration for individual behaviors and general rules [15].

Carrying new and rich information, these aforementioned big data have served different prediction tasks in all the domains of

* Corresponding author.
E-mail address: lingli@cueb.edu.cn (L. Li).

https://doi.org/10.1016/j.bdr.2021.100289
2214-5796/

society, nature and biology, and have enhanced the prediction accuracy [6]. In social prediction, big data have helped understand the general rules of human behaviors (e.g., personality traits [16] and emotions [17]), market factors (e.g., demands [18] and sales [19]), social events (e.g., election [20]), transportation (e.g., traffic flow [21] and traffic congestion [22]), etc. In the natural domain, big data have made great contributions to improving the prediction for weather factors (e.g., precipitation [23] and temperature [24]), environmental factors (including normal factors (e.g., air pollutants [25]) and emergencies (particularly natural disasters [26])), engineering issues (e.g., machine useful life [27] and machine faults [28]), material properties (e.g., material composition [29], performance [30] and structure [31]), etc. In biological forecasting, the research hotspots using big data were biomedicine (with hot issues being clinical detection [32] and epidemic propagation [33]), biotechnology (particularly genetic engineering [34] and protein engineering [35]) and animal and plant science (with prevailing prediction targets of animal abundance [36] and plant resources [37]).

Big data, with the 5V character, have the following two unique advantages in the field of forecasting research over traditional data. On the one hand, with the inherent characteristics of large *volume*, *variety* and *velocity*, big data can finely address the data limitations of small sample size, simple data type and out-of-date information from which the forecasting research using traditional data (particularly survey data) often suffered [38]. On the other hand, with the natures of *value* and *veracity*, big data appear a predictive power in terms of carrying valuable and actual information, which greatly enriched the inputs of forecasting models [39]. Nevertheless, different types of big data have their own distinctive information, unique characteristics and different formats, thereby dominating different forecasting tasks requiring different data analysis techniques to extract the hidden predictive knowledge, and being suitable for a variety of forecasting models. Therefore, a comprehensive review on full-scale types of big data is extremely admirable, detailing what (regarding the specific data types and sources), where (for the prediction hotspots) and how (for the analysis technologies and forecasting models used) big data were applied to forecasting research.

However, a systematic literature review on full-scale types of big data in forecasting research is still lacking. As listed in Table S1, the existing related reviews regarding big data in forecasting were confined to: a type of big data, such as smart meter data [40] and bio-medical data [41–43]; a type of data analysis techniques, such as feature selection [44]; a type of prediction targets, such as energy systems [45], social problems [46,47], natural issues [48,49] and bio-medical issues [43]; and/or a type of forecasting models, such as machine learning tools [50] and deep learning models [48]. Against this background, this paper attempts to fill in this literature gap to conduct a comprehensive review covering full-scale types of big data in forecasting research (and all the associated prediction tasks and forecasting models thereof).

The main aim of this paper is to present a comprehensive review on full-scale types of big data in forecasting research, detailing what, where and how big data improved prediction. Relative to the existing literature, the major contributions of this survey can be summarized into the following two aspects:

(1) This paper is the first attempt to review full-scale types of big data in forecasting research (and all the associated prediction tasks, analysis techniques and forecasting models thereof), whereas the others were limited to a specific type of big data, prediction tasks, analysis techniques and/or forecasting models;

(2) For each type of big data, an overall review is provided detailing the specific types and sources (i.e., what big data to use), forecasting hotspots (where the big data dominated) and analysis and forecasting models (how to use the big data in prediction improvement), as well as the associated major findings and future directions.

The remainder of the paper is organized as follows: Section 2 analyzes the general development of big data-based forecasting research based on descriptive and scientometric statistics and provides the general framework of the review. Sections 3–5 detail, for UGC data, device data and log data, respectively, what (regarding the specific data types and sources), where (the research hotspots), and how (the analysis and forecasting models used) big data were used in prediction, and then outlines the major findings and further directions. Section 6 concludes the review and offers helpful insights into the further prospects of this promising research, i.e., big data-based forecasting.

2. General development

This section introduces descriptive and scientometric statistics to investigate the general development of big data-based forecasting research. Section 2.1 describes the literature collection. Sections 2.2–2.5 analyze the general growth, publication sources, spatial distribution and research hotspots, respectively. Section 2.6 formulates the general framework of applying big data to forecasting research.

2.1. Databases

The existing articles using big data in prediction/forecasting are retrieved from authoritative academic databases, such as Emerald Insight, Google Scholar, Journals Online, SAGE Springer, Science Direct, Web of Science and Wiley Online Library (listed in alphabetic order). To obtain a full-scale result, we apply the keywords of *big data* AND *forecast** OR *predict**, where the symbol * represents a derivative for a term, such as “forecasting” for “forecast” and “prediction” for “predict”. Notably, the two words “prediction” and “forecasting” are technically different: prediction is a more general term; in contrast, forecasting is a specific case of prediction, which is more dependent on data and thus is more accurate [51,52]. Thus, the introduction of big data (with the distinct 5V characters [3] and at higher frequency) can finely provide new, higher-frequency information to forecasting, thereby enhancing the forecasting accuracy and temporal resolution [53]. However, most studies used the two terms interchangeably despite their differences, thus we have to do the same for a full-scale review. The period of literature is set to up to the end of 2019. Only regular articles in English are included in the analysis, while book reviews, viewpoints, research notes, reports and short communications are excluded. The downloaded literature is carefully double-checked for relevance to the forecasting research using big data. Finally, a total of 5,463 papers are selected in the review.

2.2. General growth

Fig. 1 illustrates the number of the articles on big data-based forecasting research published each year, and three important findings can be concluded. First, scholars started introducing big data into forecasting research in 2004, i.e., 3 years after the original concept of big data (i.e., 3V raised in 2001) [2]. This result indicates that big data-based forecasting research has a relatively long history, starting at the early stage of the big data era. Second, the annual numbers of published articles exhibited a generally increasing trend from 2004 to 2019, reflecting a growing interest in this promising field and revealing the effectiveness of big data in forecasting research. Third, there existed an obvious growth point in 2014, in which the annual number of publications doubled and since then it has developed rapidly with a yearly increase of over 150 papers until 2019. This great growth was supported by the

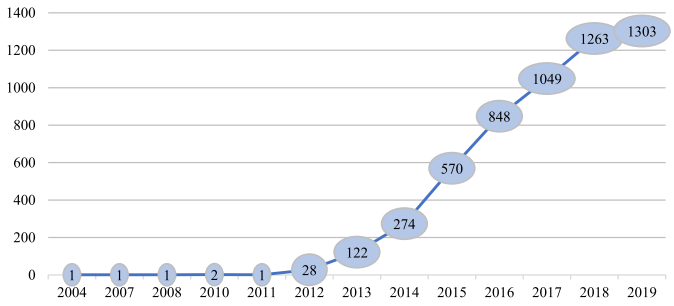

Fig. 1. Temporal trend of the publications.

second generation of blockchain (a decentralized, distributed and public digital ledger) in 2014 (blockchain 2.0), greatly enlarging data storage capacity and providing a spacious platform for openly sharing big data [54].

2.3. Publication sources

As shown in Fig. 2, the major sources were journal papers (covering a total of 1,913 journals and accounting for 79.48% of the total articles), and 20.52% of publications were conference papers. The top five leading journals were *International Journal of Forecasting* (representing 6.73% of the total articles), *IEEE Access* (4.10%), *Energies* (1.94%), *PLoS One* (1.25%) and *Future Generation Computer Systems* (1.14%). Besides comprehensive journals (e.g., *PLoS One* and *Applied Sciences*), the leading journals are majorly associated with the main factors in big data-based forecasting, such as big data (e.g., *Big Data Research*) and the associated characteristics (e.g., *Complexity*), data sources (e.g., *Sensors*) and analysis technologies (e.g., *Cluster Computing*), forecasting research (e.g., *International Journal of Forecasting*), and computer science (e.g., *IEEE Access*, *Future Generation Computer Systems* and *Computers & Electrical Engineering*). In addition, some energy-related journals (e.g., *Applied Energy* and *Energies*) have also made a large contribution to the publications, which implies that energy systems might be a forecasting hotspot using big data.

2.4. Spatial distribution

Fig. 3 and Fig. S1 illustrate the contribution network of countries or regions to big data-based forecasting research, to capture the associated leading countries, cooperation dynamics and influential relationship. For multi-country papers, a common paper is computed once for each of the associated cooperative countries in counting the countries' respective publications and cooperations [55]. The top five influential contributors (in terms of the number of publications, represented by the sizes of nodes) include the USA (with 1,595 articles, representing 29.20% of the total articles), China (1,561 articles, 28.57%), India (422 articles, 7.72%), England (323 articles, 5.91%) and South Korea (252 articles, 4.61%). Regarding cooperative time (represented by the color of links), the earliest international collaboration occurred in 2008 (between Spain and France), and the cooperation began flourishing across the top five leading countries in 2010. The cooperative relationship was the strongest between the USA and China (with 288 collaborations, accounting for 3.93% of the total cooperative articles; represented by thickness), the USA and England (101, 1.38%), China and Australia (58, 0.79%), and China and Germany (54, 0.74%). In comparison, Iceland and El Salvador had few international cooperations, which subsequently limited their contribution to this research (representing 0.02% of the total articles for both).

2.5. Hot keywords

Fig. 4 and Fig. S2 show the keyword co-occurrence network from a timezone view, in which the hot keywords co-concurring more than 2 times are listed by frequency (represented by the front size of keywords and the dimension of circles). There were few hot keywords from 2004 to 2010 whereas a rapid increase in emerging hot keywords since 2011, reflecting increasingly great attention to the application of big data to forecasting in recent years. Not surprisingly, the searching keywords that are used to collect articles in this review are identified as hot keywords (listed in Table S2): "big data", "prediction" and "forecasting" as well as their derivative terms (such as "big data analytics", "predictive analytics" and "predictive modeling"). Furthermore, the other hot keywords reveal the prevailing types of big data, forecasting hotspots, analy-

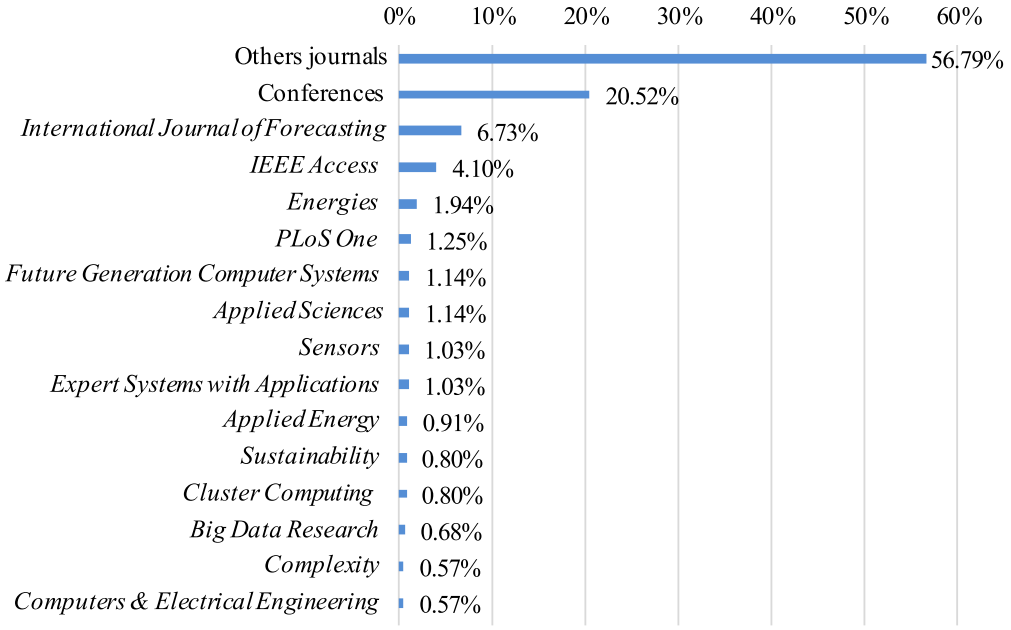

Fig. 2. Distribution of the publication sources.

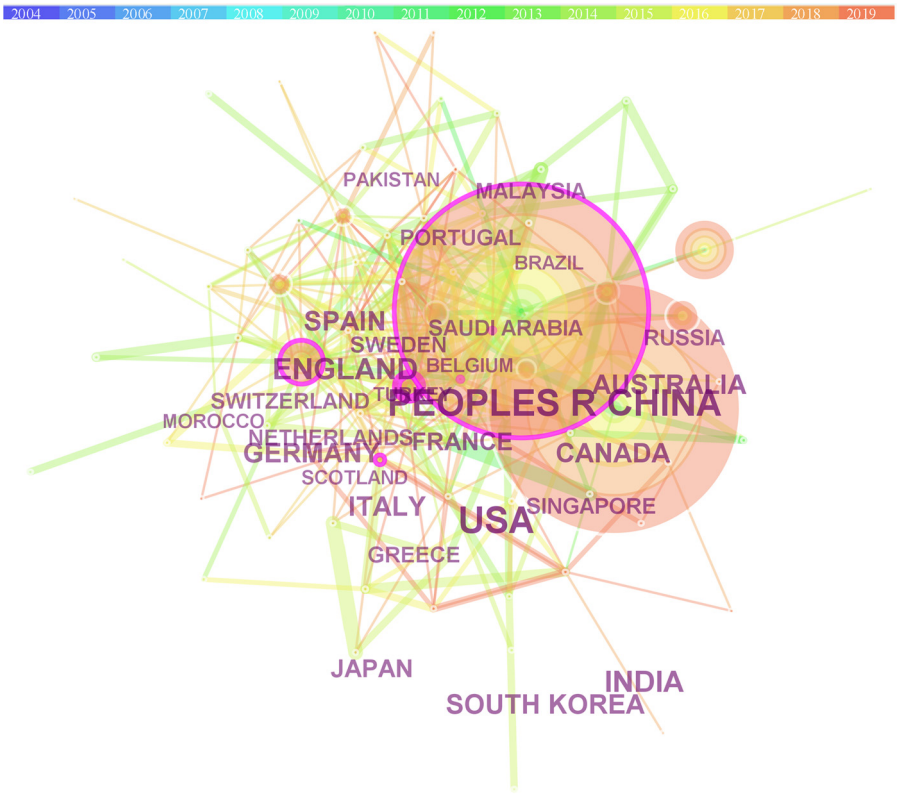

Fig. 3. Contribution network of countries or regions. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

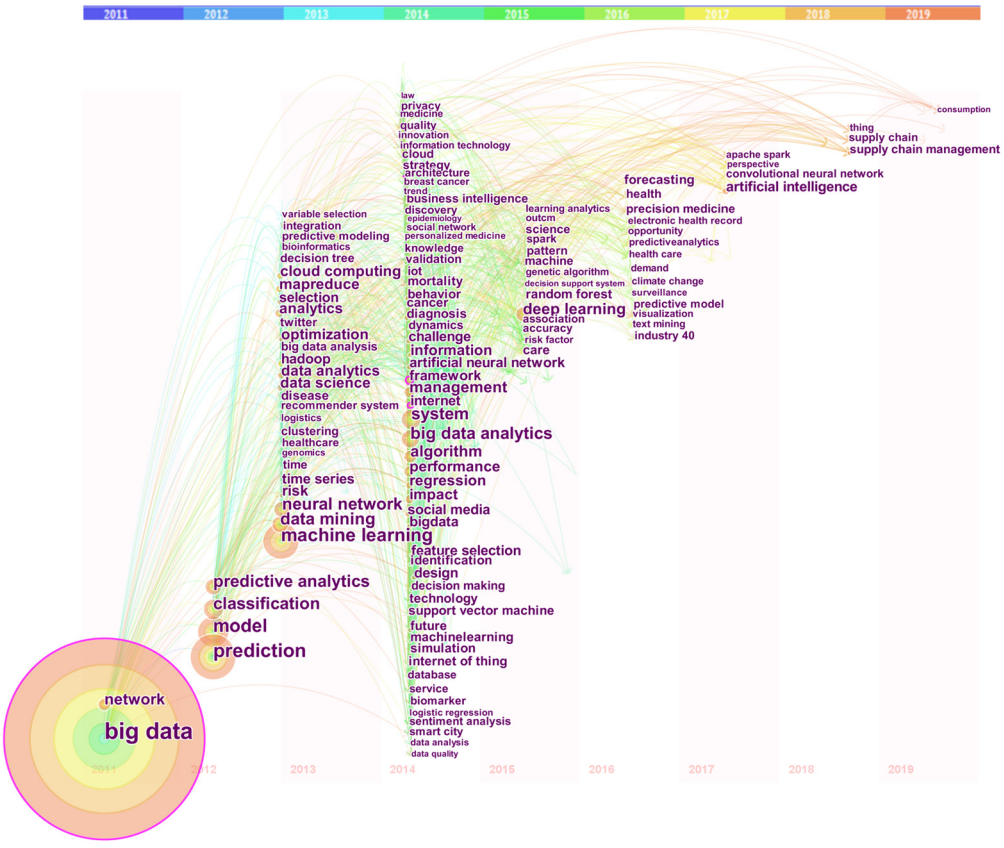

Fig. 4. Co-occurrence network of keywords.

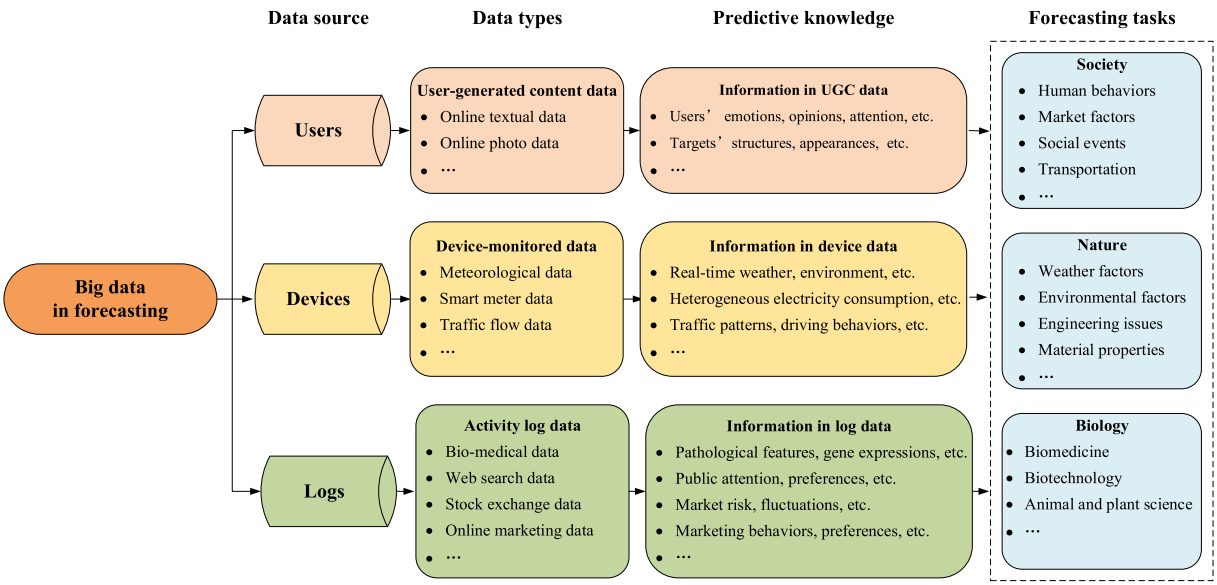

Fig. 5. Big data in forecasting research.

sis technologies and forecasting models in big data-based forecasting research.

2.6. General framework

According to existing related literature, different types of big data have been and will still be introduced to forecasting research. There exist many classification methods for data or big data, and here we group big data based on source or origin (regardless of other features, such as data format or dataset file) [56]. By source, the big data in forecasting research generally fall into three main categories, i.e., UGC data (generated by Internet users), device data (measured by monitoring devices) and log data (recording activities), with each type providing its own distinctive new knowledge to the prediction for diverse social, natural and biological factors or issues, as illustrated in Fig. 5.

According to Fig. 5, the big data in forecasting research can be generally categorized into three major types by source: UGC data (generated by the users on social media or other web platforms), including online textual data, online photo data, etc.; device data (monitored by devices), including meteorological data, smart meter data, traffic flow data, etc.; log data (recording activities or operations), including bio-medical data, web search data, stock exchange data, online marketing data, etc. Different types of big data have provided different new, rich knowledge to prediction: for example, users' emotions, opinions and attention toward prediction target-related events or issues from UGC data [16,57,58]; the sensor-level, real-time dynamics of weather environment (e.g., real-time temperature and wind speed) [25], electricity consumption [39] and traffic behaviors [59] from device data; medical insights (e.g., regarding pathological features and gene expressions) [60,61], public attention and preferences (in prediction-related activities) [62], market dynamics [63] and marketing behaviors and preferences [64] from log data. This informative knowledge facilitated various challenging forecasting tasks: in the domain of society, the forecasting hotspots were the dynamics of human behaviors [13], market factors [19], social events [65] and transportation [66]; in the nature, big data primarily served the prediction for weather factors [23], environmental factors [25], engineering issues [28] and material properties [29]; in the biology, popular research fields were biomedicine [67], biotechnology [35] and animal and plant science [37]. Interestingly, a variety of standalone applications have been designed to help users perform specific tasks, and the related big data (even in different output files) can be similarly classified by data source, i.e., the users who employ the applications or the targets whose actions are recorded by the applications [56,68].

Generally, three major steps are taken in big data-based forecasting research, i.e., data collection (to collect big data from the associated sources), data processing (to preprocess and represent the big data and to extract the predictive knowledge) and prediction improvement (by incorporating the extracted predictive knowledge into forecasting models as important inputs), as illustrated in Fig. 6. In data processing, a set of analysis technologies have been employed to extract the insightful information hidden in big data through the following three sub-steps: (1) data preprocessing, with the main aim to identify and treat unrelated, duplicated, missing and abnormal values [65,69]; (2) data representation, to transform big data into a structured format [70] and into a comparable unit [71]; (3) feature selection, to extract useful information through word vector selection for textual data [72], image feature selection for photo data[13], informative feature extraction from massive features [39] and relationship exploration to find prediction target-related features [25]. In prediction improvement, the predictive knowledge extracted in Step 2 is put into forecasting models as important inputs. The prevailing forecasting models can fall into three categories: statistical models [73,74], artificial intelligences (AIs) [65,75] and hybrid models [25].

Fig. 7 shows the distributions of the articles using different big data, for different forecasting hotspots and via different forecasting models in big data-based forecasting research. For data types, device data (accounting for 54.14% of the total studies) and log data (31.92%) dominated the big data in forecasting research, of which meteorological data (representing 31.53% of the articles using device data, which is mainly attributable to large attention to weather forecasting globally) and bio-medical data (accumulating at a large speed and on a large scale [76,77] and arousing wide attention due to the high importance of bio-medicine research [77]; representing 43.04% of the articles using log data) made the greatest contribution. The top research hotspots lied in the society (accounting for 49.70% of the total articles), given that big data can effectively reflect individual consumer behaviors [8,78]. For forecasting models, AI models dominated the forecasting models

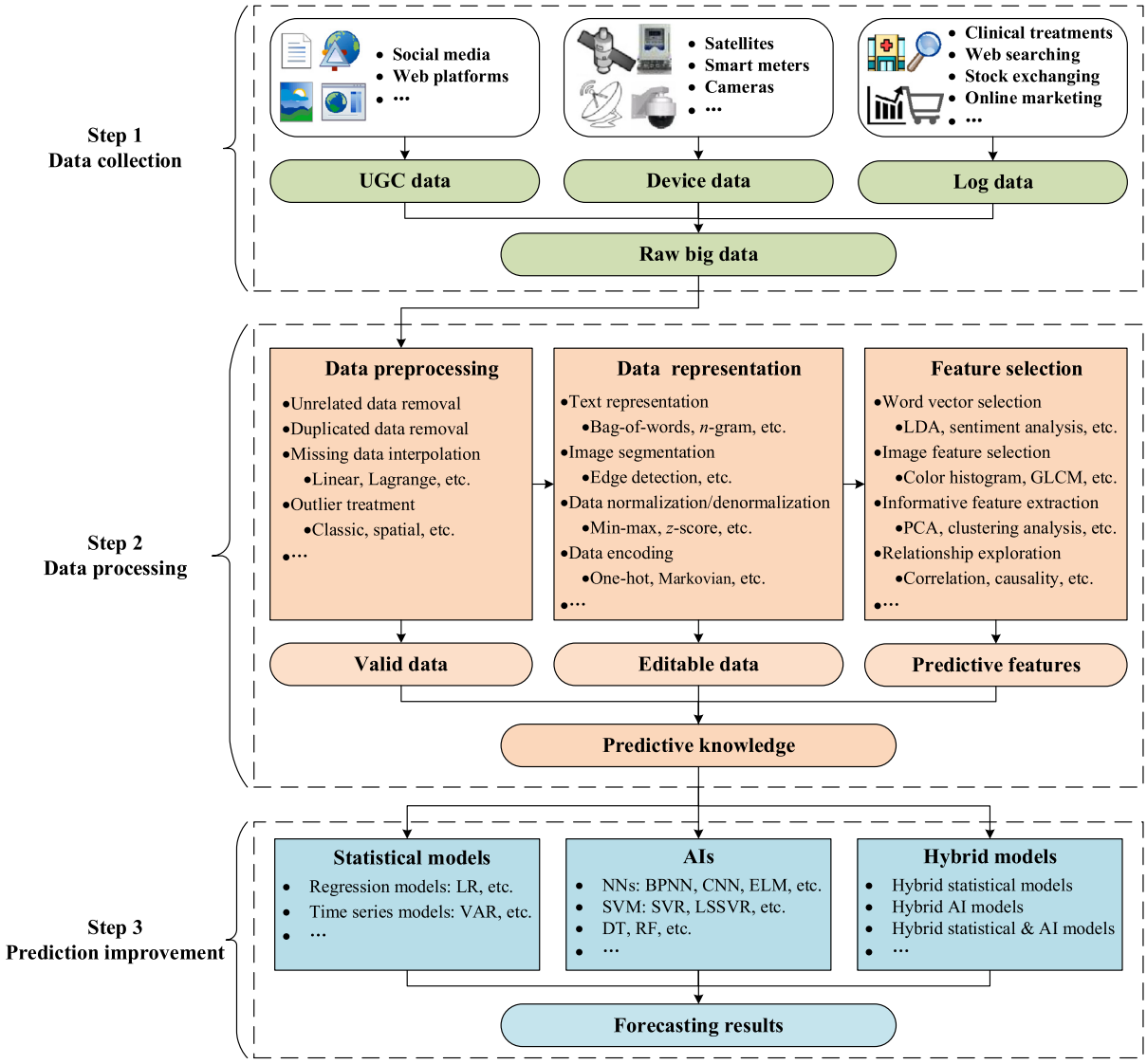

Fig. 6. General framework of big data-based forecasting research.

using big data (representing 60.94% of the total articles), particularly neural networks (NNs) and support vector machines (SVM) (representing 41.02% and 18.42%, respectively, of the AI-based articles), mainly due to their powerful ability to address nonstationary, nonlinearity and complexity (which are the main characteristics of almost datasets in big data infrastructure [25,79–81].

Sections 3–5, for UGC data, device data and log data, respectively, elaborate on what (regarding the specific data types and sources thereof), where (regarding the forecasting hotspots) and how (regarding the analysis technologies and forecasting models used) big data improved forecasting research.

3. UGC data

In the digital era, a rich mine of interesting information can be posted and shared by the users on social media and other web platforms, forming UGC data [8] and bringing new knowledge (e.g., public opinions [65], emotions [16] and attention to prediction target-related events or issues [57]) to forecasting research. This section details what (in Section 3.1), where (Section 3.2) and how (Section 3.3 and 3.4) UGC data improved prediction.

3.1. Data types and data sources

As a prevailing type of big data, UGC data, including online textual data (13.13%) and online photo data (0.81%), have helped to improve forecasting research (accounting for 13.94% of the total articles). The online textual data in forecasting came from a variety of social media (e.g., Twitter [75], Facebook [82] and Sina Weibo [83]) and other websites (e.g., the shopping websites of Amazon [18] and Taobao [84]), where the users can express and share their individual experiences, viewpoints, reviews or other information [8]. The main types of online textual data were comments/reviews [18], posts [85], news [57], reports [86] and blogs [75]. Apart from textual data, online photo data, which were majorly derived from the Google image search [13], Twitter [87], Facebook [88] and Twitch [89], have also been used in forecasting.

3.2. Forecasting hotspots

Online textual data and online photo data have been extensively applied to forecasting research and have served different forecasting tasks, as illustrated in Fig. 8.

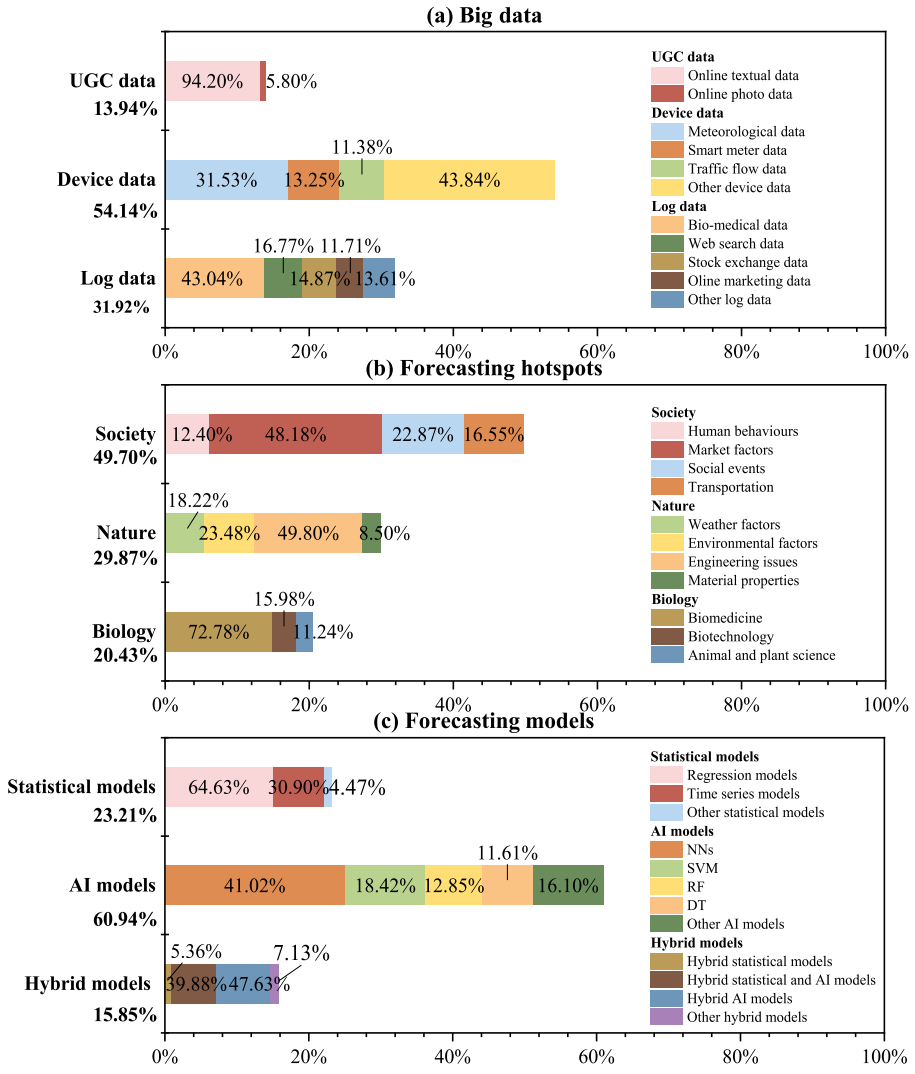

Fig. 7. Distributions of big data (**a**), forecasting hotspots (**b**) and forecasting models (**c**) in big data-based forecasting research.

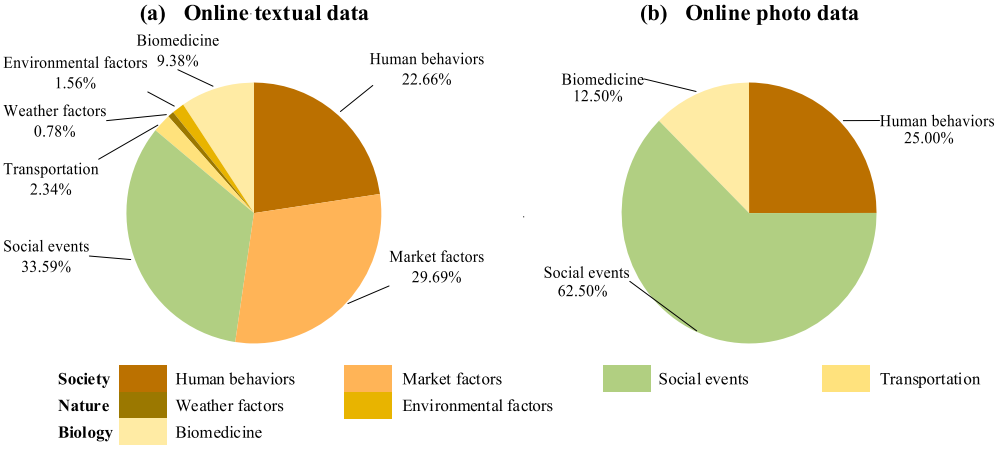

Fig. 8. Distributions of UGC data, i.e., online textual data (**a**) and online photo data (**b**), in prediction.

As shown in Fig. 8, online textual data and online photo data have both been majorly applied to social prediction (representing 88.28% and 87.50% of the articles using online texts and photos, respectively), particularly for social events (33.59% and 62.50%) and human behaviors (22.66% and 25.00%), and biological prediction for biomedicine (9.38% and 12.50%). Interestingly, online textual data absolutely dominated the UGC data used in forecasting (presenting 94.20% of the UGC-based papers), thereby having a far

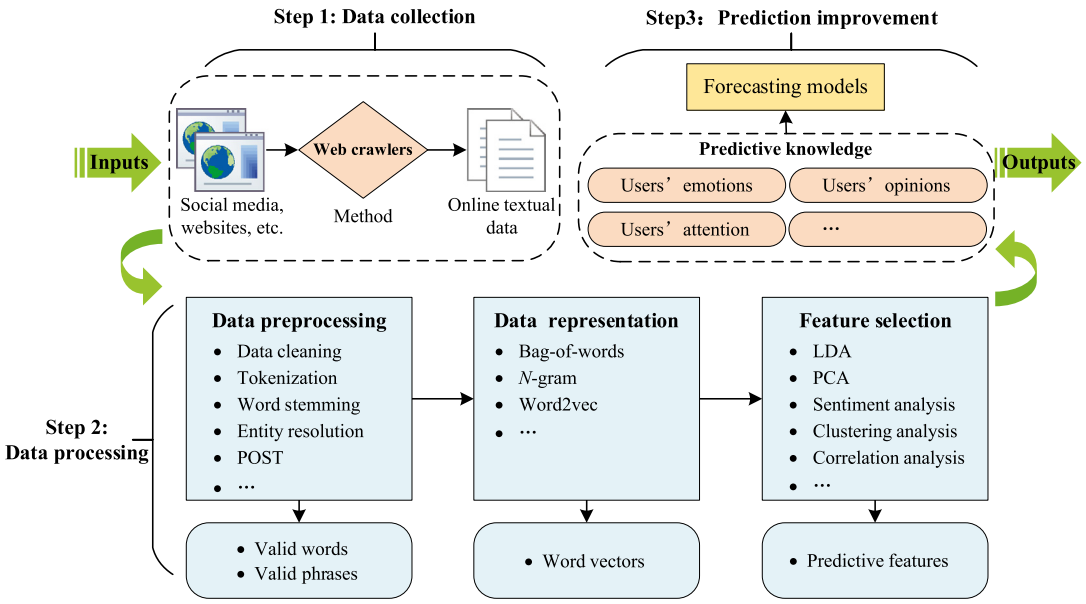

Fig. 9. General framework of online textual data-based forecasting research.

wider range of forecasting research (see Fig. 8-(a)). In the society, besides the above issues, the forecasting hotspots also covered market factors (representing 29.69% of the articles using online texts, particularly product sales [19]) and transportation (2.34%, particularly traffic conditions [21]). In the nature, environmental factors (1.56%, particularly $PM_{2.5}$ concentrations [58]) and weather factors (0.78%, particularly temperature [90]) were also forecasting hotspots that used online textual data.

3.3. Analysis technologies

A variety of analysis technologies have been proposed and introduced to collect UGC data from diverse social media or web platforms (in Step 1), to process the data through the three sub-steps of data preprocessing, data representation and feature selection for extracting the hidden knowledge (Step 2), and to put the extracted new knowledge into forecasting models (Step 3).

3.3.1. Online textual data

Fig. 9 describes the general procedure of online textual data-based forecasting, together with the associated text mining techniques.

(1) Data collection

In the first step of data collection, diverse web crawlers [91] have been designed and implemented to withdraw online textual data from the associated social media or other websites. In particular, a web crawler (namely, robot, spider and worm) is actually a set of programs automating the tasks of visiting web pages and downloading the corresponding textual contents [92]. The common programming languages of web crawlers include Python [91] and Java [19].

(2) Data processing

In the second step of data processing, three sub-steps were generally taken to analyze the collected online textual data and obtain the hidden knowledge: data preprocessing, data representation and feature selection.

In data preprocessing, the operations of data cleaning, tokenization, word stemming, entity resolution and part-of-speech tagging (POST) were popularly conducted in online textual data-based forecasting research, to filter valuable words or phrases from massive online texts.

- *Data cleaning*, which tends to identify and remove invalid words, such as misspelling [72], stop words [65], punctuation marks [72], tabs [93], incomplete words [91], non-target language and low frequency words [16,91] from the texts, leaving valid information.
- *Tokenization*, aiming to split sentences into a stream of words or phrases (namely, tokens) for further analysis [16].
- *Word stemming*, to recognize the root of words with equal (or similar) meaning and to reduce the number of words, in order to save data processing time and memory space [16].
- *Entity resolution* (or duplicate identification and record linkage), to identify and group the records (in terms of words or phrases in text) referring to the same real-world entity [94–96].
- *POST*, determining the POS tag for each word, leaving the words with the POS tags of noun, verb and adjective and removing unimportant words with other tags [72].

Through these operations, original online textual data can be transformed into a substantially reduced number of valid, meaningful, important words or phrases.

In data representation, a series of analysis methods (e.g., bags-of-words, *n*-gram model and Word2vec) have been employed to transform the extracted words into word vectors (i.e., document matrices) or other structural formats, with elements commonly representing the weights (e.g., in terms of frequencies) of the associated words.

- *Bag-of-words*, an effective object categorization method, which quantizes words (or strings) into word vectors and calculates the frequency of each word [70] based on different algorithms (e.g., term frequency/inverse document frequency (TF-IDF) [70]).
- *N-gram model*, measuring the co-occurring frequencies of each *n*-gram (defined as *n* adjacent words in a sentence or string, commonly referred to as a shingle), and selecting the shingles of high co-occurring frequencies for vector transformation [97, 98].

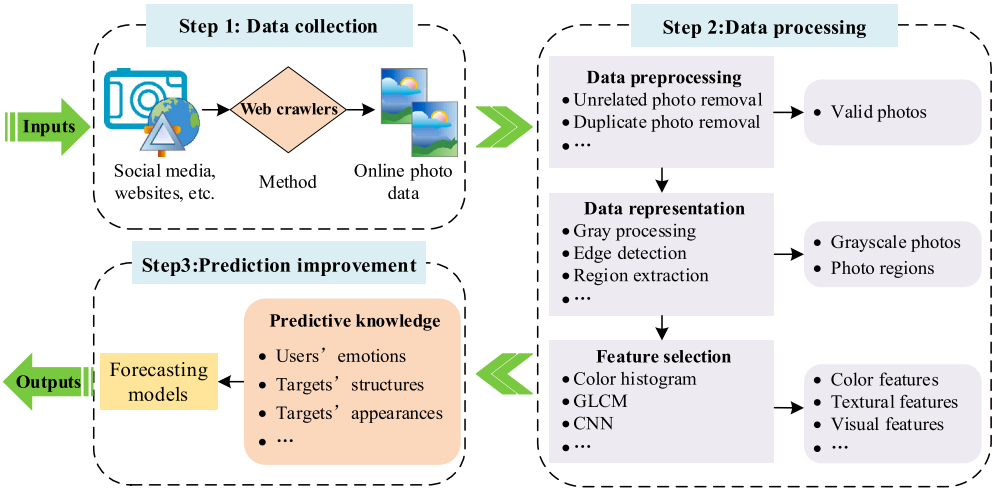

Fig. 10. General framework of online photo data-based forecasting research.

– *Word2vec*, a deep learning model proposed by Google in 2013, aiming at clustering words with similar meaning and mapping them into a vector based on the occurrence frequencies of words [99].

Based on these technologies, online textual data can be represented in word vectors (or document matrices) for later quantitative analysis.

In feature selection, a series of statistical analyses, e.g., latent Dirichlet allocation (LDA), principal components analysis (PCA), sentiment analysis, clustering analysis and correlation analysis, have been employed to analyze the features or variables (in terms of word vectors or document matrices constructed above) and extract the predictive ones (with insightful knowledge and herewith predictive power for the prediction targets).

– *LDA*, tending to cluster similar features (in terms of co-occurring word vectors or document matrices) into one topic and then to capture the latent topic structure (i.e., semantics) as predictive features that can be used as model inputs for further prediction improvement [72].
– *PCA*, mapping the vector space into a feature space to extract the principal components as predictive features that have low dimensions but carry most data information [10].
– *Sentiment analysis*, determining the polarity of words or phrases (positive, negative or neutral emotion) and extracting the emotional features that are closely related to the prediction target [86].
– *Clustering analysis*, such as naive Bayes classifier [17], grouping similar features into a group and detecting the predictive features closely related to the prediction target.
– *Correlation analysis*, such as Pearson correlation analysis, measuring the relationship between variables and the prediction target and identifying the predictive features with a high correlation [16].

Through these statistical analyses, the predictive features for prediction targets can be selected from the word vectors, which can be used as an important model input in prediction.

(3) Prediction improvement

In the third step of prediction improvement, the extracted predictive features (carrying valuable knowledge and predictive power thereof) were put into existing forecasting models as important inputs beyond traditional data. On the one hand, the basic information of online texts is considered to be predictive, particularly users' features (e.g., user ID) [91], temporal features (like timestamps) [88] and textual features (such as the volume of texts) [58], which jointly reflects the public attention of the related agents to the research targets [57]. On the other hand, online textual data also provided new insightful knowledge that traditional data cannot provide, such as public emotions (being positive, negative or neutral toward the research target [16,17]) and opinions (e.g., supporting [62,65]), which can be captured by sentiment analysis from blog [65], post [85] and comment data [18].

3.3.2. Online photo data

Fig. 10 provides the general framework of applying online photo data to forecasting.

(1) Data collection

Similar to online text data, web crawlers in Python and MATLAB were used to automatically download online photo data from social media or websites [13].

(2) Data processing

In data preprocessing, aiming to filter valid photos from massive metadata, the antipole-tree algorithm [13] was popularly employed in online photo data-based forecasting research, to find and delete unrelated photos and duplicate photos.

In data representation, gray processing, edge detection and region extraction have been widely used to transform photos to editable formats (e.g., grayscale images) and then to divide them into photo regions.

– *Gray processing*, with the aim to quantitatively describe the pixels of a color photo in terms of numbers, for example, mapping light intensity into a continuous scale between 0 (black) and 255 (white) [100]. Based on this technique, color photos can be transformed to grayscale photos.
– *Edge detection*, exploring abrupt pixel changes as the edges to segment a photo into disjoint regions [101], with popular techniques of Sobel edge detector [89] and 2D Gabor wavelet transform in online photo data-based forecasting [100].
– *Region extraction*, to partition a photo into regions based on pixel similarity [102], with the dominant methods of region growing-and-merging method [103], region splitting-and-

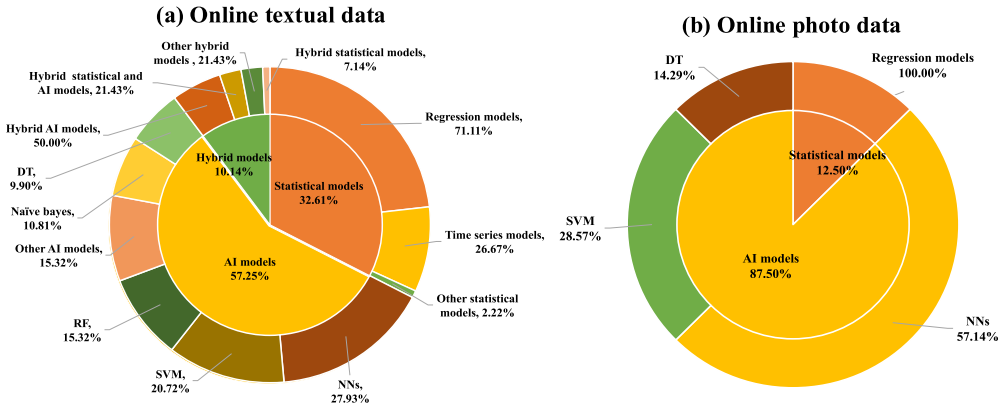

Fig. 11. Distributions of forecasting models in forecasting research using UGC data, i.e., online textual data (**a**) and online photo data (**b**).

merging method [104] and watershed transformation method [105].

In feature selection, photo regions were characterized and clustered to photo features and then selected to predictive features, via popular techniques of color histogram, gray level co-occurrence matrix (GLCM) and convolutional neural networks (CNN).

- *Color histogram*, which groups photo regions to color features based on color distributions, from which predictive features can be selected to provide prediction-related information (e.g., emotions reflected by color brightness [13]).

- *GLCM*, clustering photo regions based on the grayscale difference between pixels (measured via entropy, correlation, contrast, homogeneity, etc.) and capturing the textural features (e.g., photo structures based on the spatial arrangement of regions [106]) to describe the structures of the prediction target [107].

- *CNN*, a powerful deep learning method used here to characterize photo regions by emulating the optic nerve behavior of living organisms and to excavate the visual features of prediction targets (such as facial features [108]).

(3) Prediction improvement

In prediction improvement, the photo features obtained above—in particular, the color, structure and visual features of a photo, carrying insightful knowledge and herewith predictive power [13]—have been introduced as key predictors into forecasting models, substantially enhancing the prediction accuracy for users' emotion [13], users' characteristics [107], face recognition [108], etc. The basic information of photos, e.g., uploader's ID [109], uploading location [13], uploading time [107] and the total volume of related photos [87], has effectively helped understand the behaviors of the users uploading the photos. Besides, photo features can effectively reflect users' emotions (by photo colors [13]) and vividly describe the structures (by textural features [106]) and appearances or looks of prediction targets (by visual features [106]).

3.4. Forecasting models

Various forecasting models have been employed to incorporate the insightful features hidden in online textual and photo data into prediction, as shown in Fig. 11. Among them, statistical models and AI models dominated UGC data-based research (jointly representing 89.86% and 100.00% of online textual data- and online photo data-based articles, respectively), and hybrid models (combining statistical and/or AI models) have recently been introduced to the research using online textual data (Fig. 11-(a)) but not yet for online photo data (Fig. 11-(b)).

For statistical models (jointly representing 31.25% of UGC-based articles), diverse regression analyses [65] and time series models [83] have been employed in UGC data-based forecasting. Among the two subtypes, regression analyses appeared far more popular (accounting for 71.11% and 100% of online textual and photo data-based articles, respectively, using statistical models): for example, linear regression (LR) [82] and logistic regression for online textual data [65] and LR models for online photo data [88]. In comparison, time series models were used only for online textual data (26.67% of online text-based articles using statistical models), with typical models of autoregressive integrated moving average (ARIMA) particularly for news data [8] and blog data [83].

Due to the complexity of UGC data, AI models appeared to be much more powerful in UGC data-based prediction (representing 59.03% of UGC data-based articles), particularly NNs (29.66% of UGC data-based articles using AIs) [18], SVM (21.19%) [70], random forests (RF, an extension of decision trees (DT)) (14.41%) [85] and DT (10.17%) [75]. For both online textual data- and online photo data-based forecasting, NNs (representing 27.93% and 57.14%, respectively, of the associated articles using AIs), SVM (20.72% and 28.57%) and DT (9.90% and 14.29%) were popularly used. Furthermore, online textual data-based research had a far larger variety in forecasting models, covering many more AI algorithms, such as RF (accounting for 15.32% of online textual data-based articles using AIs) and naive Bayes (10.81%).

Hybrid models, even as newcomers in UGC data-based prediction, had a boom in recent years (representing 9.72% of UGC data-based articles by the end of 2019). Diverse hybrid models have been designed, especially by combining two or more AI models (50.00% of UGC data-based articles using hybrid models) [99], statistical and AI models (21.43%) [110] and statistical models (7.14%) [111], with the main principle of taking advantage of a model to address the weaknesses of others [99]. However, such a promising approach, which has shown its superiority in prediction [99], has only been applied to the research using online textual data (accounting for 10.14% of online text-based studies) [99] and even not yet for those using online photo data.

3.5. Major findings and future directions

The boom in social media has largely promoted the application of UGC data in prediction, particularly online textual data and online photo data. However, there is still ample room to improve such promising research. For data types, the UGC data used in existing forecasting studies mostly include online texts and photos in relatively simple, easy-to-use formats. Regarding other UGC data, audio data and video data (e.g., from Netflix, YouTube, Vine and

Facebook) [112–114] have also been used in predictive analytics, particularly in marketing for customer behaviors [112], video incomes [115], video popularity [114], etc. In data analytics, there is only a slight additional complexity in processing such data types. Specifically, audio data can be run through a speech-to-text process, and the loss of information due to that translation is not usually significant for predictive purposes [116]. Similarly, video data can be processed into less complicated data for further analyses, mainly including: (1) web log data, regarding the upload time of video, number of clicks/views/loops, etc. [117,118]; (2) textual data, e.g., the subtitles and comments of video [113,114]; (3) photo data in video frames, for capturing image features like color, objective and image quality [113,114]; (4) audio data extracted from video [114] and then textual data through the speech-to-text process. In prediction analytics, AI models (particularly support vector regression (SVR), CNN and long short-term memory (LSTM)) prevailed in the prediction models using such UGC data [113,119,120]. However, such data types have been relatively less used in prediction, compared with other UGC data. The possible hidden reasons might lie in the massive amount, large size and therewith lots of information contained, which require huge storage room and processing power [121,122].

For forecasting hotspots, UGC data have sufficiently been found to be insightful for social prediction (representing 88.23% of UGC data-based articles, particularly for human behaviors and social events). However, such useful data were otherwise less popular in the domains of biology and nature (11.77% jointly) and even have not been applied to some important forecasting tasks, such as biotechnology and animal and plant science in the biology, engineering issues and material properties in the nature [123]. The hidden reason might be that the UGC data used in prediction were mainly from social media (such as Twitter, Facebook and Sina Weibo), which effectively reflected individual information and helpfully captured public opinions, emotions and attention (highly associated with social prediction). In comparison, there are relatively fewer platforms sharing the data for biological and natural prediction, most of which are still in the control of industries. Therefore, the construction of enterprise public information disclosure systems [124] might be a promising way to promote UGC data application in biological and natural prediction.

For analysis technologies, a variety of data mining techniques (with popular tools of tokenization and word stemming techniques for texts, and gray processing and GLCM method for photos) have been applied to extract predictive knowledge from UGC data. However, it was still a challenge to extract key features from UGC data, avoiding too many predictors in prediction. Effective big data analysis technologies can also be introduced to capture other interesting features, such as latent aspect rating analysis for textual data to estimate the ratings of individual opinions and the relative emphasis on each aspect [125] and semantic-based image retrieval to mine information hidden in the semantic annotations of photo data [126].

For forecasting models, AI models were the most prevalent in existing UGC data-based forecasting research (accounting for 59.03% of the associated articles), while hybrid models, a rising star in prediction, were relatively few (9.72%). However, hybrid models, finely combining models to take advantage of a model to address the weakness of others, have powerful strengths in complex system prediction [99], particularly in the context of big data. Therefore, such a promising type is strongly recommended to capture the complex relationship between UGC data and prediction targets, particularly in online photo data-based prediction where hybrid models have not yet been introduced.

4. Device data

With the help of the IoT technology, diverse devices, monitors or sensors have been used to automatically track individual prediction-related targets and their specific changes at a monitor level and in real time. This generated device data [127,128] and provided monitor-level, real-time knowledge about weather environment [25], individual electricity consumption [39], traffic patterns [59], etc., which greatly enhanced the spatiotemporal resolution of forecasting models.

4.1. Data types and data sources

Device data have fully shown the feasibility and superiority in the existing forecasting research (representing 54.14% of the total studies), including meteorological data (17.07%), smart meter data (7.17%), traffic flow data (6.16%) and other device data (23.74%).

In particular, the meteorological data in forecasting (representing 31.53% of device data-related forecasting research), monitored by weather station sensors [23], meteorological radars [129] and satellites [130], mainly included upper air data [25], surface data [130], radiation data [131], radar data [129] and satellite data [130].

The smart meter data used in prediction (13.25%) were measured by smart meters to track residential and commercial electricity consumption and loads [71]. Such data were provided by open databases [132] and power system operators, such as State Grid Corporation of China (SGCC) [133], Independent System Operator of New England (ISO NE) [14]) and Green Button (GB) of New Zealand [39].

Traffic flow data (11.38%) were monitored by the cameras and sensors installed on roads [134], to track dynamic traffic conditions [66]. In addition, other device data (jointly 43.84%), such as product testing data (10.82%) [27], equipment operation data (10.81%) [28], GPS data (8.02%) [11], mobile phone data (7.84%) [135], air quality monitoring data (5.04%) [25], and automatic identification system (AIS) data (1.31%) [136], have also been introduced to improve prediction.

4.2. Forecasting hotspots

Fig. 12 reveals that device data-based forecasting hotspots were social prediction (representing 25.00-96.72% of the articles using any of the four main types of device data), especially for market factors (4.92-51.43%) [137] and social events (1.19-9.36%) [135]; natural prediction (3.28-63.10%), particularly for engineering issues (3.28-47.14%) [28].

Different device data, offering distinctive information, dominated different research domains. For meteorological data (see Fig. 12-(a)), apart from the above hotspots, the forecasting focuses also included transportation (accounting for 7.14% of meteorological data-based studies, particularly traffic safety [11]) and human behaviors (1.19%, human activities [138]) in the society domain; weather factors (25.00%, temperature [24]), environmental factors (24.40%, air pollutions [25]) and material properties (0.60%, material performance [139]) in the nature; animal and plant science (9.52%, vegetation attributes [37]) and biomedicine (2.38%, epidemic propagation [33]) in the biology. Traffic flow data (Fig. 12-(c)) also served the prediction for transportation (representing 81.97% of traffic flow data-based articles, particularly traffic conditions [21]) and human behaviors (4.92%, human driver intent [140]) in the society. Other device data of product testing data, equipment operation data, GPS data, mobile phone data, air quality monitoring data and AIS data (Fig. 12-(d)) also improved the prediction for transportation (representing 11.91% of the articles

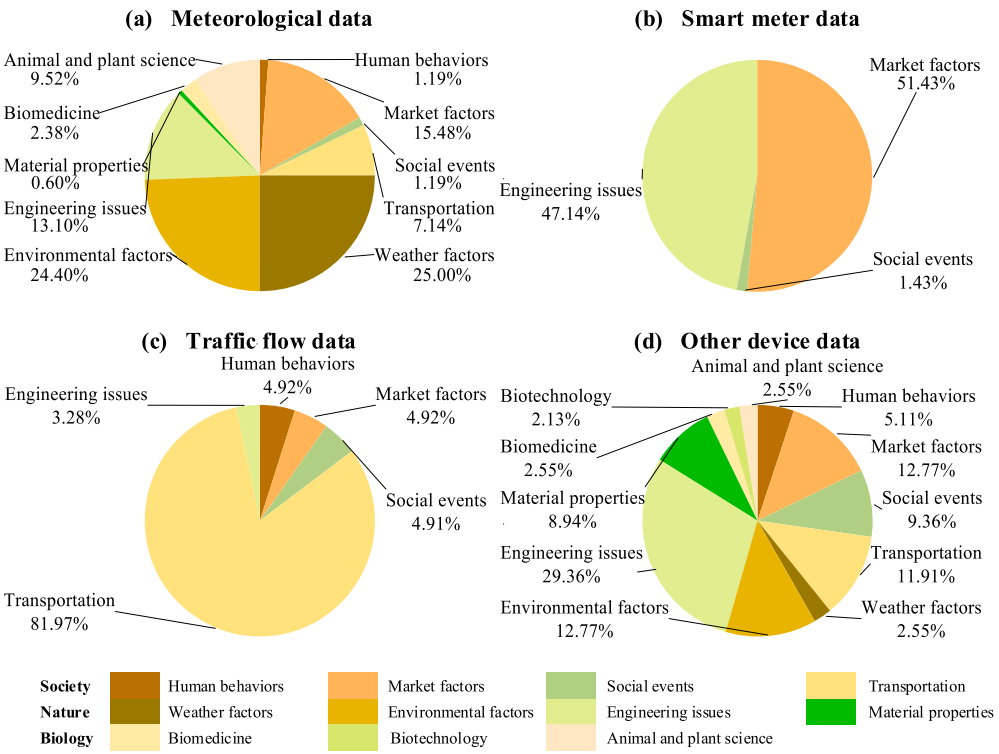

Fig. 12. Distributions of device data, i.e., meteorological data (**a**), smart meter data (**b**), traffic flow data (**c**) and other device data (**d**) in prediction.

using all other device data, particularly traffic risk [11]) and human behaviors (5.11%, human destination [141]) in the society; environmental factors (12.77%, air quality [69]), material properties (8.94%, material tensile strength [30]) and weather factors (2.55%, precipitations [142]) in the nature; biomedicine (2.55%, epidemic propagations [143]), animal and plant science (2.55%, crop yields [144]) and biotechnology (2.13%, bio-pharmaceuticals [145]) in the biology.

4.3. Analysis technologies

Figs. 13–15 provide the specific frameworks of meteorological data-, smart meter data- and traffic flow data-based forecasting research, respectively, together with the corresponding analysis techniques. Sections 4.3.1–4.3.4 elaborate on how to collect and analyze each type of device data and extract the hidden valuable knowledge for prediction.

4.3.1. Meteorological data

As before, the three steps of data collection, data processing and prediction improvement were taken in meteorological data-based prediction, as illustrated in Fig. 13.

(1) Data collection

Massive sensors or monitors have been installed and applied globally to monitor sensor-level, real-time meteorological conditions, forming meteorological data. The meteorological data in forecasting research mainly came from weather stations (with the Hoshiarpur and Patiala stations [131] and Dumdum meteorological station [146] being typical examples), meteorological radars (e.g., windfinding radars [147] and weather radars [129]) and meteorological satellites (e.g., Landsat [130] and FENGYUN meteorological satellite [148]). Meteorological data can be divided into two subtypes in distinct formats: numerical data, recorded in text, extensible markup language (XML), hypertext markup language (HTML), etc. [25]; image and video data [130], where video data were commonly converted to image data at a given frequency for data analysis.

(2) Data processing

In data preprocessing, the analysis tools of linear interpolation [149] and missForest algorithm [131] have popularly been used to treat missing data in numerical meteorological data, while the radiometric calibration method, geometric correction method and atmospheric correction method [130] have extensively been employed to calibrate image meteorological data.

- *Linear interpolation*, which simply interpolates missing meteorological data by linearly connecting adjacent measurements [149].
- *MissForest algorithm*, an iterative imputation method to fill missing values by training an RF model [131].
- *Radiometric calibration method*, converting the original digital number of meteorological images to radiant luminance values, to avoid sensor errors and obtain high-quality meteorological data [130].
- *Geometric correction method*, to normalize the geographic coordinates of satellite images by shifting, rotating, scaling and warping the meteorological images [150].
- *Atmospheric correction method*, converting the original digital number of meteorological images to surface reflectance values, to remove the effects of the atmosphere [130].

Data representation was conducted to obtain dimensionless datasets and valid image regions from numerical and image meteorological data, respectively. To normalize numerical meteorological data, effective tools were min-max normalization [151] and null mean normalization [152].

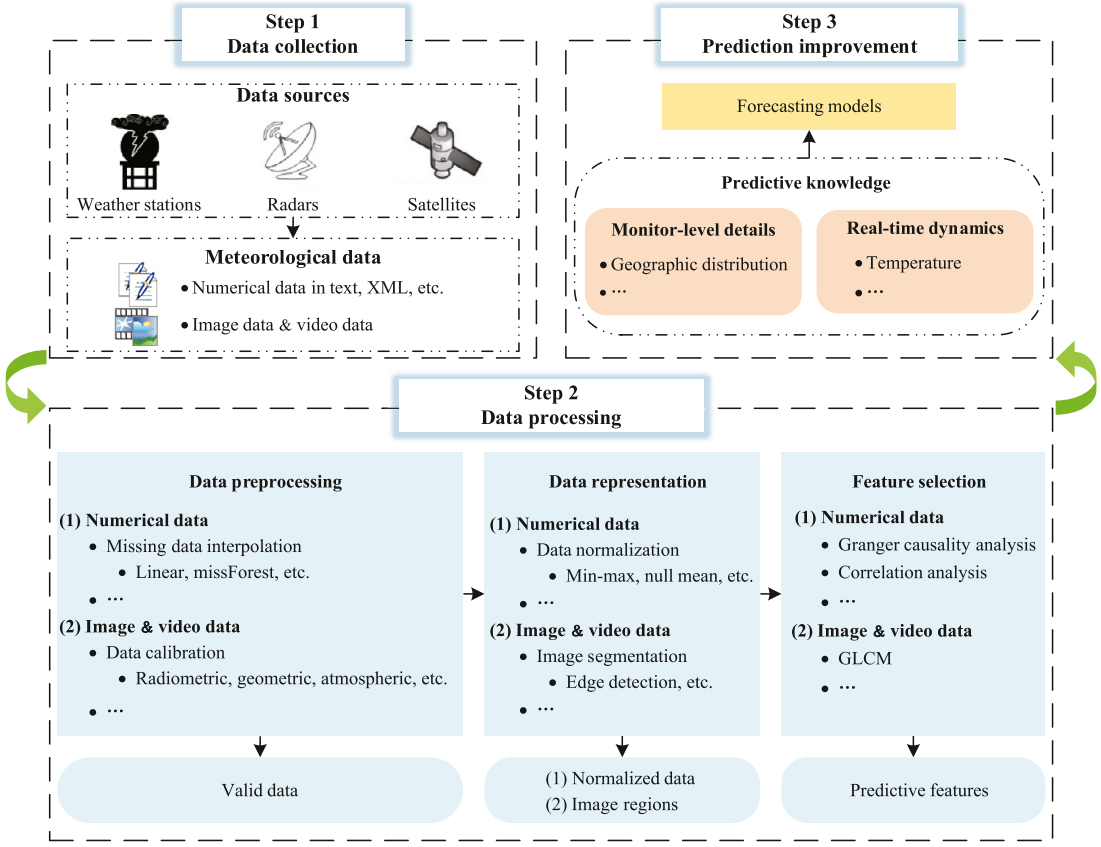

Fig. 13. General framework of meteorological data-based forecasting research.

- *Min-max normalization*, shifting the minimum and maximum scores of datasets to −1 and 1, respectively, to eliminate data dimensions [151].
- *Null mean normalization*, to normalize meteorological measurements into a common range based on the sample mean [152].

For image meteorological data, effective image processing techniques, e.g., edge detection and region extraction (see details in Section 3.3.2), have been popularly employed to split an image into image regions [153].

In feature selection, relationship exploration analyses (e.g., Granger causality analysis and correlation analysis) and image feature extraction techniques (e.g., GLCM; delated in Section 3.3.2) have been introduced to obtain predictive features from numerical and image meteorological data, respectively. *Granger causality analysis* and *correlation analysis* identify the predictive features in a significant causal and linear relation, respectively, to the prediction targets [25].

(3) Prediction improvement

In prediction improvement, the valuable predictive features (derived from both numerical and image meteorological data) were put as important inputs into forecasting models. Meteorological data described not only the monitor-level details of prediction targets (such as geographic distribution [25]) but also their real-time dynamics of temperature [131], humidity [25,131], wind speed [25], precipitation [146], etc., which largely enhanced the prediction accuracy for weather or weather-related issues or events.

4.3.2. Smart meter data

(1) Data collection

Smart meters have extensively been installed in residential and commercial buildings, recording individual electricity consumption and loads [71,154]. The smart meter data used in prediction mainly fell into two sub-types by source: open-access data from open databases (such as the Dataport dataset [132]) and private data belonging to power system operators (such as SGCC, ISO NE and GB [14,133]).

(2) Data processing

In data preprocessing, to detect and process invalid data (e.g., nulls, zeros or outliers), the effective techniques of the Lagrange interpolation formula [133], classic outlier approach [155] and spatial outlier approach [155] have been widely used in smart meter data-based forecasting.

- *Lagrange interpolation formula*, interpolating invalid smart meter measurements for a complete and valid dataset, based on a polynomial regression [133].
- *Classic outlier approach*, aiming to detect and remove outliers that significantly deviate from the general trend in probability, distance, deviation or density [155].
- *Spatial outlier approach*, to delete outliers with a significant deviation from the general trends in both spatial and non-spatial attributes [155].

In data representation, *z*-score normalization [71] and min-max normalization [156] have popularly been applied to remove the differences in scale and measurement units for consistency, based on the arithmetic mean and standard deviations of smart meter data.

In feature selection, various smart meter measurements (for different attributes, at different frequencies and/or with different

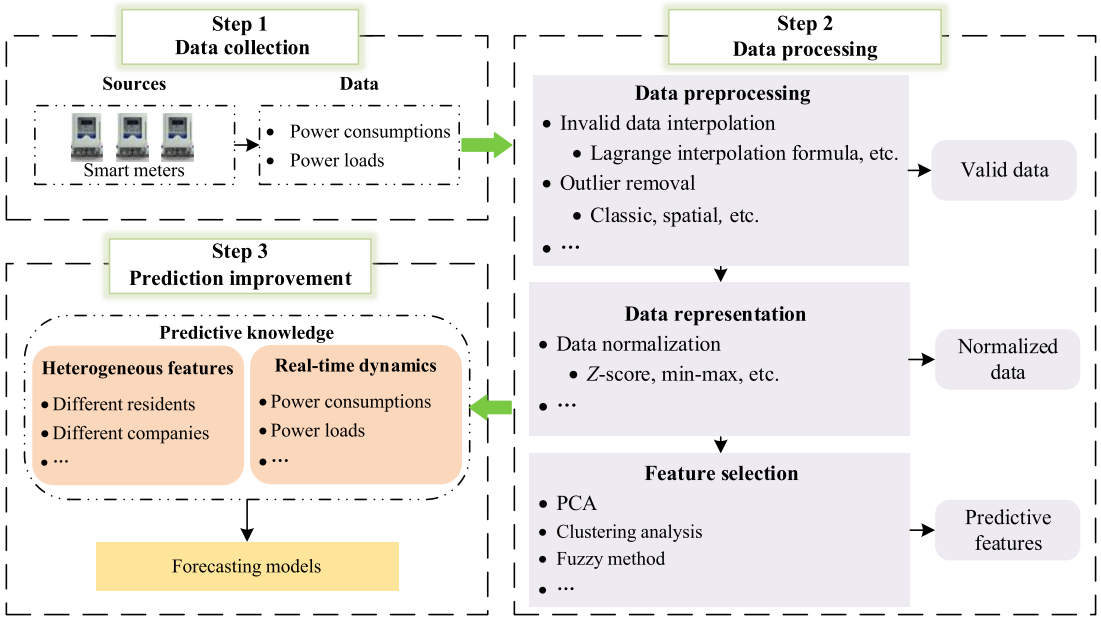

Fig. 14. General framework of smart meter data-based forecasting research.

time lags), together with other related factors, were grouped and selected into predictive features, with typical methods of PCA, clustering analysis and fuzzy method.

- *PCA*, mapping smart meter data and other related factors into principal components and selecting the ones that account for the most predictive information [39].
- *Clustering analysis*, such as the *k*-means algorithm, to cluster variables (including various smart meter measurements) into groups as features and then find the predictive features in close relation to the prediction targets [39].
- *Fuzzy method*, using fuzzy evaluation and decision methods to evaluate and rank candidate features (including various smart meter measurements) and select the most informative ones [155].

(3) Prediction improvement

The unique predictive knowledge in smart meter data was the real-time dynamics of electricity consumption [39] and electricity loads [133,154], as well as the associated heterogeneous features across different residents and companies, which have been used as effective predictors to largely enhance the spatiotemporal resolution and accuracy of existing forecasting models.

4.3.3. Traffic flow data

(1) Data collection

The traffic flow data in prediction were measured by various cameras and sensors installed on roads, including automatic traffic recorders [59], automatic number plate recognition sensors [21], motorway incident detection and automatic signaling loop sensors [21], traffic monitoring unit loop sensors [21], etc. Traffic flow data have two main data types: numerical data, such as time series [157]; image data, involving images [66] and videos that were commonly converted to image frames for data analysis [134].

(2) Data processing

In data preprocessing, data cleaning and data conversion were commonly conducted.

- For *numerical data*, data cleaning was performed to detect and discard abnormal, duplicate and unrelated data, to gain valid data from massive traffic flow data [157].
- For *image data*, data conversion was conducted to transform traffic flow videos to image frames at regular intervals, for further data analysis [66].

In data representation, normalized datasets or editable image regions were extracted from numerical or image traffic flow data, respectively, via different techniques.

- For *numerical data*, min-max normalization [157] and *z*-score normalization [158] were useful methods to remove data dimensions.
- For *image data*, threshold segmentation was popularly used to split an image frame into parts by setting pixel thresholds, for example, to distinguish vehicles from the background [159].

In feature selection, predictive features or image features that are related to forecasting targets were selected for numerical and image data, respectively.

- For *numerical data*, effective statistical analyses, such as clustering analysis and correlation analysis [137], have been used to group traffic flow data into features and then identify the predictive feature with a high correlation with prediction targets.
- For *image data*, predictive image features (e.g., vehicle position and vehicle speed) were calculated by comparing adjacent image frames [66,134].

(3) Prediction improvement

Based on the basic information in traffic flow data (e.g., locations of sensors [134] and timestamps [66]), the real-time patterns of vehicle position [134], vehicle speed [134] and traffic flow [137]) can be well tracked and have been used as important factors describing driving behaviors [134] and transportation system dynamics [137], thereby enhancing the associated prediction accuracy.

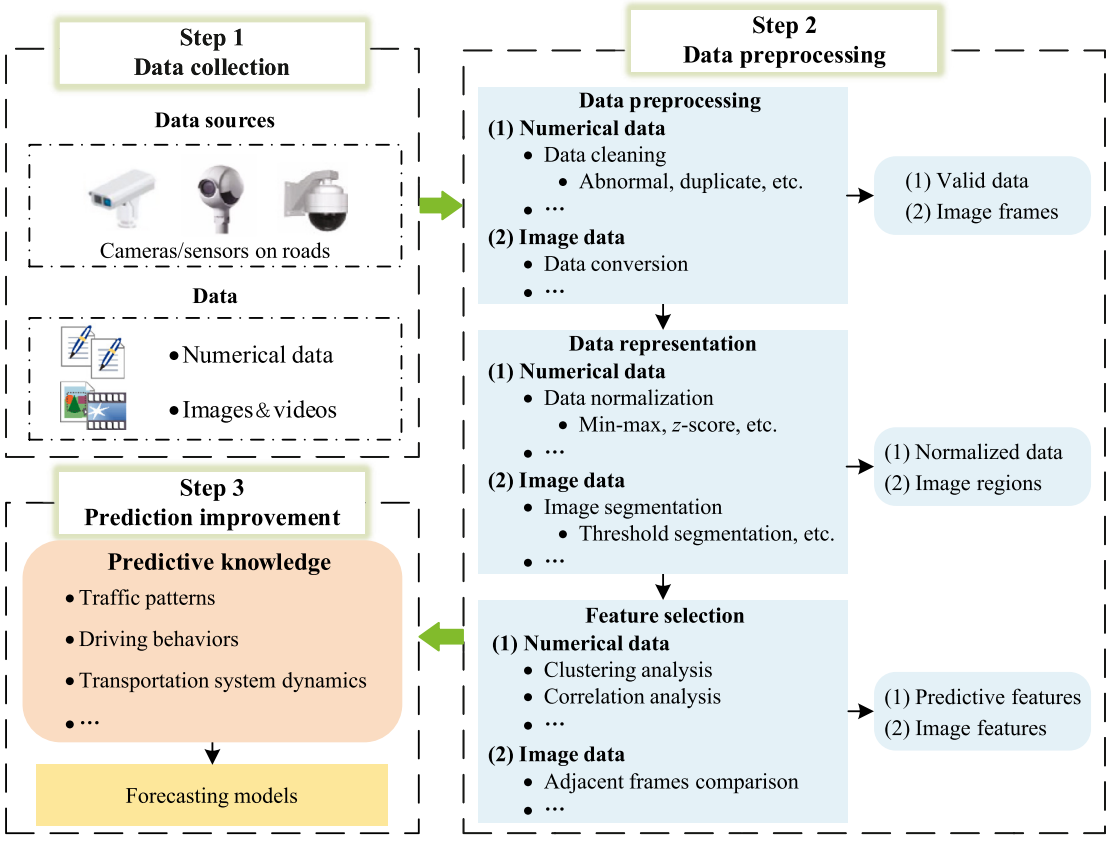

Fig. 15. General framework of traffic flow data-based forecasting research.

4.3.4. Other device data

In addition to the aforementioned data, a variety of other device data have also been employed to improve prediction research, including product testing data, equipment operation data, GPS data, mobile phone data, air quality monitoring data and AIS data.

(1) Product testing data

In *data collection*, the product testing data in prediction research were generated from the examination of the quality and property of products via a variety of sensors or devices, such as fault detection devices, metal detection devices and mechanical test sensors [27,30]. In *data processing*, the sub-step of data cleaning was conducted to delete missing and incorrect data from the raw datasets [30]; data representation conducted data normalization with typical method of min-max normalization [27]; feature selection has introduced emerging data mining methods (such as deep neural networks (DNN) [160]) to adaptively capture predictive knowledge, apart from statistical analysis (e.g., correlation analysis [161]). In *prediction improvement*, the predictive features regarding the performance [160], composition [29] and degradation pattern of products [27] were popularly used as effective predictors into forecasting models.

(2) Equipment operation data

In *data collection*, intelligent sensors have been installed on the targeted equipment to record the real-time status of performance and operation, forming equipment operation data [28]. In *data processing*, statistical imputation methods have been applied to address missing data [162]; effective methods, such as PCA and Fisher criterion method, have been employed to extract informative features [28]. In *prediction improvement*, the extracted predictive features helpfully captured the general patterns in equipment functions and faults, thereby greatly improving the associated predictions [28].

(3) GPS data

In *data collection*, the GPS data in prediction were gathered in two main ways: GPS loggers carried by vehicles (including taxies and bicycles) [11] and GPS-enabled mobile applications installed in smartphones [163]. In *data processing*, data preprocessing was performed to transform raw data to editable formats (e.g., vectors, matrices or tensors); data representation focused on data normalization, especially via min-max normalization [164]; feature selection popularly employed clustering analysis to group features [165], correlation analysis to find predictive features [166] and machine learning (such as CNN) to capture the trajectory features of targets [11]. In *prediction improvement*, the predictive features extracted from the basic information in GPS data (e.g., timestamps [11], latitude and longitude [166]) can help capture the dynamic trajectory [11] and moving behaviors of prediction targets [164] and were introduced as important inputs into existing forecasting models.

(4) Mobile phone data

In *data collection*, the mobile phone data in forecasting were emitted and received by telecommunication base stations through radio waves, and stored and provided by mobile network operators [167]. In *data processing*, popular data analyses included data characteristic explorations (e.g., augmented Dickey-Fuller (ADF) test) to find valid signals of stationarity and clustering analyses (e.g., *k*-means clustering algorithm) to aggregate similar features into a group as a predictive feature [135]. In *prediction improvement*, the

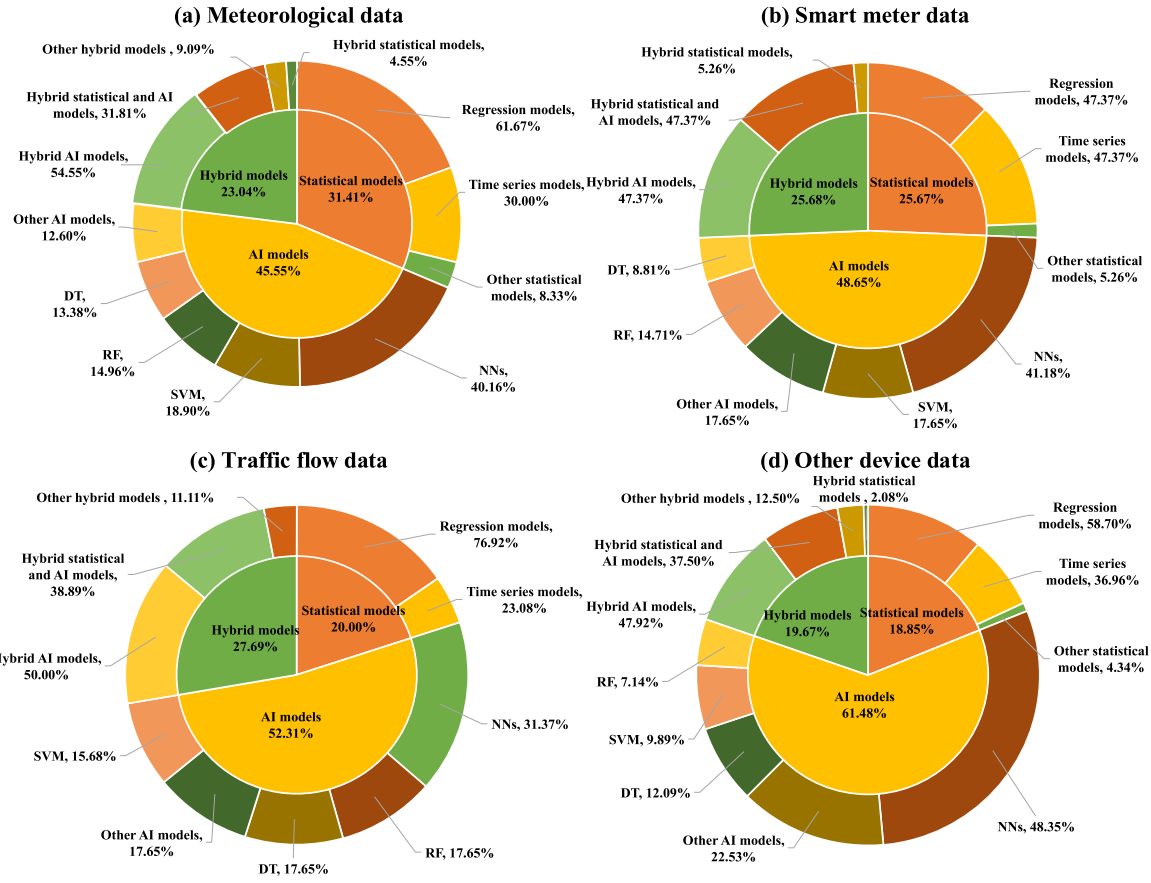

Fig. 16. Distributions of forecasting models in forecasting research using device data, i.e., meteorological data (**a**), smart meter data (**b**), traffic flow data (**c**) and other device data (**d**).

extracted predictive features, which finely reflect the spatiotemporal distributions of mobile users, calls, short messages, etc. [167], have appeared to have predictive power in the related prediction tasks.

(5) Air quality monitoring data

In *data collection*, the air quality monitoring data in prediction were mainly obtained from national air quality monitoring stations [168], which measure and record the real-time concentrations of pollutants (e.g., $PM_{2.5}$, NO_X and SO_2) and evaluate the air quality indices. In *data processing*, data preprocessing was conducted to address missing values and outliers [69]; data representation focused on data normalization, particularly via min-max normalization [169] and batch normalization [170]; feature selection introduced relationship exploration analyses (e.g., correlation analysis, cointegration analysis and causality analysis [25]) to examine the predictive power of features, machine learning algorithms (e.g., CNN [170]) to select predictive features in a data-driven, self-adaptive way, and time series analyses (e.g., multivariate empirical mode decomposition [25] and wavelet decomposition [168]) to mine the hidden factors at different timescales (or frequencies). In *prediction improvement*, the pollution concentrations of $PM_{2.5}$, NOx, SO_2, etc. have generally been considered as effective factors in air quality monitoring data-based prediction [69], as well as the associated hidden factors at different timescales [25].

(6) AIS data

In *data collection*, the AIS data in prediction were collected and offered by onboard transceivers and terrestrial and satellite base stations [171]. In *data processing*, data preprocessing was performed to obtain complete, valid AIS datasets using interpolation methods, such as the Hermite interpolation method [136]; data representation focused on data normalization, e.g., based on batch normalization technology [171]; feature selection extensively introduced clustering analysis to group features [136]. In *prediction improvement*, based on the basic information in AIS data (e.g., vessel's name, length, breadth, longitude, latitude, navigation time, speed and heading [136]), the extracted predictive features finely reflected trajectory dynamics and kinematic behaviors of prediction targets [171], largely improving the associated predictions.

4.4. Forecasting models

According to Fig. 16, AI models prevailed in device data-based forecasting research (representing 54.07% of the associated articles), followed by statistical models and hybrid models (23.59% and 22.34%, respectively).

Among statistical models, regression analyses made a large contribution to device data-based prediction improvement (representing 61.06% of device data-based articles using statistical models), such as LR for meteorological data [172] and smart meter data [173] and logistic regression for traffic flow data [74]. In addition, time series models also prevailed (31.86%), with typical cases of using ARIMA for meteorological data [58] and traffic flow data [22] and autoregressive methods for smart meter data [174].

AI models dominated device data-based forecasting, of which the models prevailing in all types of device data were NNs (representing 42.14% of device data-based articles using AIs) [160], SVM (13.97%) [153], DT (12.72%) [37] and RF (11.47%) [59]. In addition, some other AI models appeared powerful for a particular data type,

such as naive Bayes for traffic flow data (representing 8.82% of traffic flow data-based papers using AIs) [21], and Bayes network for smart meter data (representing 3.03% of smart meters data-based articles using AIs) [133].

Hybrid models, an emerging type in prediction, had a boom recently in device data-based prediction. The hybrid models were designed mainly by combining AI models (representing 50.39% of device data-based articles using hybrid models) [146] and statistical and AI models (37.21%) [156], which have extensively been used in the forecasting research using any type of device data. In comparison, hybrid models based only on statistical models were far less popular (3.10%) [175] and even fewer in traffic flow data-based prediction due to the nonstationary, nonlinearity and complexity of device data that challenge the basic data assumption of statistical models.

4.5. Major findings and future directions

The vigorous development of the IoT technologies has promoted the diversity of device data in forecasting research, covering meteorological data, smart meter data, traffic flow data, product testing data, equipment operation data, GPS data, mobile phone data, air quality monitoring data and AIS data. However, many works are still needed to take full advantage of such an insightful data type. Other invaluable device data, such as Wi-Fi data [176] and Bluetooth data [177], which have clear superiorities in convenience and low cost and have already served other research tasks (e.g., route tracking and planning [176,177]), can also be widely introduced to facilitate and improve forecasting research.

For forecasting hotspots, device data have played an important role in natural prediction (accounting for 51.90% of device data-based articles, particularly for engineering issues and environmental factors) and social prediction (40.94%, particularly for market factors and transportation). However, device data have attracted much less attention in biological predictions (7.16%), particularly for biomedicine and biotechnology. To improve such important prediction tasks, promoting health-related portable devices (e.g., wearable sports bracelet, recording heart rates and sleeping time of individuals) and carefully analyzing the associated insightful data is an interesting direction for future research.

For analysis technologies, to effectively process massive, real-time device data, some promising data processing methods are strongly recommended. For instance, data compression can be used to substantially accelerate data transmission and processing and reduce data storage space and computation time. This can help to improve the efficiency of data analysis and knowledge mining [178]. Moreover, to enhance the data quality of device data and detect measurement errors, comprehensively combining datasets and comparing data across different device data types and even other data types are strongly suggested in future research.

For forecasting models, AI models dominated existing device data-based forecasting research (accounting for 54.07% of the associated articles), and hybrid models (particularly those based on AIs) aroused increasingly large attention (22.34%). They adaptively capture the nonstationary, nonlinearity and complexity features in device data [179]. Notably, a promising type of hybrid models, i.e., decomposition and ensemble methods, recently emerged and greatly improved the prediction using meteorological data [180], smart meter data [181] and air quality monitoring data [25]. Based on the idea of “divide and conquer”, decomposition and ensemble models divide complicated device data into relatively simple components at different time scales, which reduces model complexity and enhances prediction accuracy. However, such a competitive type of method was scarcely used in the prediction using traffic flow data, GPS data, mobile phone data and AIS data, thereby implying a huge opportunity for prediction improvement.

5. Log data

Log data, recording the details (e.g., who, when, where, what and how) on activities (or transactions and operations), have provided insightful knowledge in prediction, such as bio-medical insights deduced from the logs on clinical treatments and research experiments [60,61], public attention from web searching [6], market dynamics from stock exchange [73] and user preferences in online/offline marketing [64], online visiting/browsing [86], etc.

5.1. Data types and data sources

Log data have well served forecasting research (representing 31.92% of the total articles), involving bio-medical data (13.74%), web search data (5.35%), stock exchange data (4.75%), online marketing data (3.74%) and other log data (jointly 4.34%).

Bio-medical data (representing 43.04% of log data-based articles) recorded the specific procedures and associated results of clinical treatments in hospitals or clinics and research experiments in labs [34,182], providing valuable insights into future disease diagnosis and medical research and improving the predictions [60,61]. The bio-medical data used in prediction mainly included clinical data (e.g., clinical notes, medical records, electronic health records (EHRs) and computed tomography (CT) images) [60] and bioinformatic data (e.g., genomic data and proteinic data) [34].

For web search data (representing 16.77% of log data-based articles), a web search for a given event or issue (namely, search keyword) would be logged and stored by the associated search engines, such as the Google engine popularly used worldwide [183], the Baidu engine in China [15] and the NAVER engine in South Korea [184]. The web search data are computed in terms of the absolute search volume for a keyword (e.g., Baidu index [15]) and relative search volume (e.g., Google trends [12] and NAVER trends [184]), which finely reflect public attention or preferences for the related events or issues.

Stock exchange data in prediction (representing 14.87% of log data-based articles) were majorly collected from finance websites (e.g., the Yahoo finance [185] and Google finance [186]) and stock markets (e.g., the NASDAQ Stock Exchange [73], New York Stock Exchange [73] and Shenzhen Stock Exchange (SZSE) [187]), recording the trade time, stock ID, price, trade size, buyer, seller, etc. of each exchange operation [188].

Online marketing data in prediction (representing 11.71% of log data-based articles) detailed marketing behaviors on a variety of online shopping platforms (e.g., Alibaba [189], Taobao [64] and Amazon [19]), regarding consumers, trade time, products, trade prices, trade volumes, etc. This rich information helped understand the marketing behaviors or preferences in the associated predictions.

In addition to the aforementioned data (totally representing 86.39% of log data-based articles), a series of log data have also served prediction, including web log data (9.18%) [190], smart card data (3.80%) [191] and highway traffic data (0.63%) [192].

5.2. Forecasting hotspots

As shown in Fig. 17, the forecasting hotspots for all the types of log data included social prediction (representing 2.21-100% of the articles using different types of log data), especially for market factors (0.74-93.62%) and social events (1.47-28.85%).

For bio-medical data (see Fig. 17-(a)), the forecasting hotpots were biological prediction, especially for biomedicine (representing 79.41% of the articles using bio-medical data, particularly clinical detection [32] and epidemic propagation [33]) and biotechnology (18.38%, gene technology [34] and protein engineering [35]). Web search data (Fig. 17-(b)), apart from the above common hotspots,

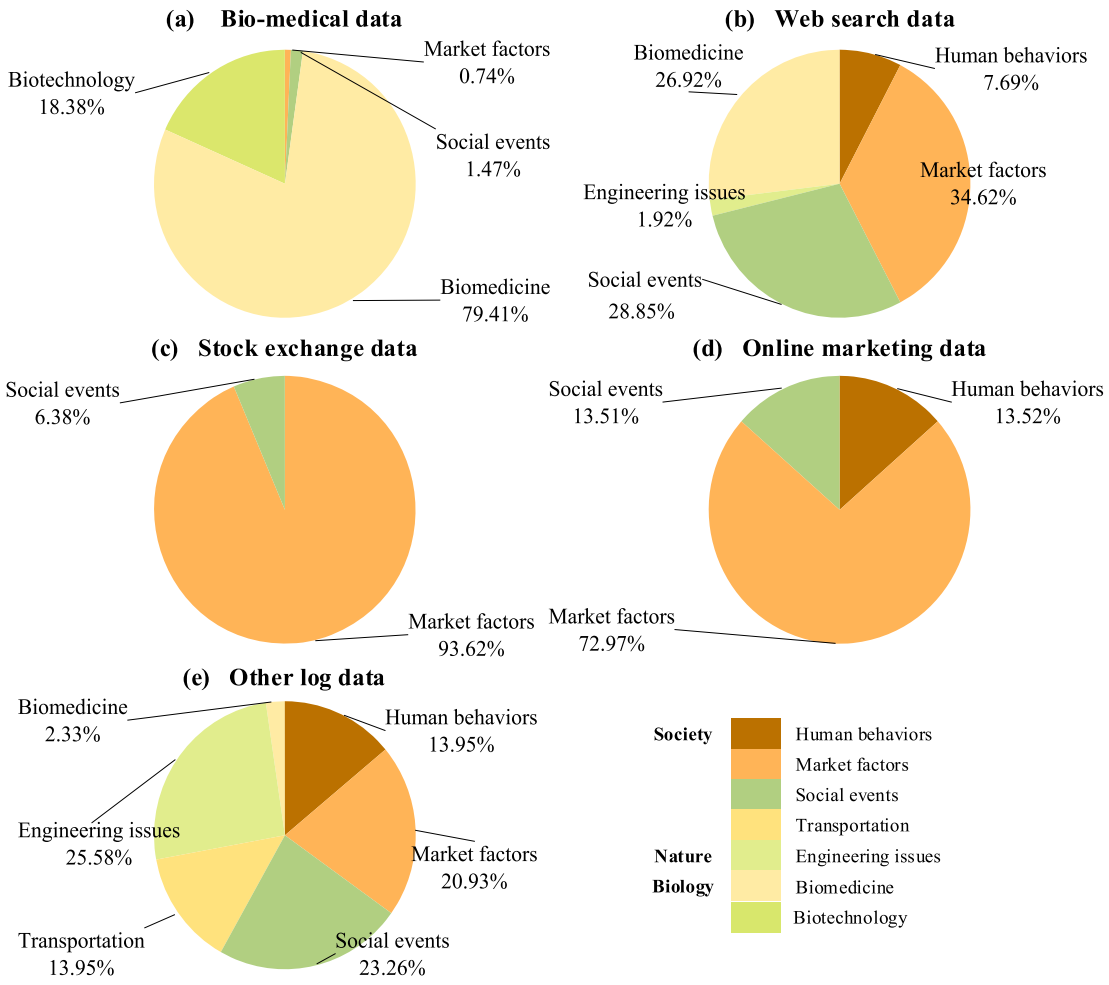

Fig. 17. Distributions of log data, i.e., bio-medical data (**a**), web search data (**b**), stock exchange data (**c**), online marketing data (**d**) and other log data (**e**) in prediction.

have also served social prediction for human behaviors (accounting for 7.69% of the articles using web search data, particularly public emotions [193]), natural prediction for engineering issues (1.92%, network engineering [194] and biological prediction for biomedicine (26.92%, clinical detection [195]). Online marketing data (Fig. 17-(d)) also facilitated other forecasting tasks of human behaviors (representing 13.52% of the articles using online marketing data, particularly customer preferences [64]). In addition, other log data of web log data, smart card data and highway traffic data (Fig. 17-(e)) also served the predictions for human behaviors (representing 13.95% of the articles using other log data, especially path choices [196]) and transportation (13.95%, travel time [192]) in the society, engineering issues (25.58%, network security engineering [197]) in the nature, and biomedicine (2.33%, clinical detection [109]) in the biology.

5.3. Analysis technologies

Figs. 18–21 describe the general framework of using bio-medical data, web search data, stock exchange data and online marketing data, respectively, in forecasting research, and Sections 5.3.1–5.3.5 detail, for each type of log data, how to collect and process the data and to select the hidden predictive knowledge for prediction improvement.

5.3.1. Bio-medical data

(1) Data collection

The bio-medical data in forecasting research were generated and collected mainly in hospitals and clinics (recording clinical treatments) [182] and laboratories (for research experiments) [34], thereby falling to clinical data [60] and bioinformatic data [34] respectively. In particular, the clinical data in prediction mainly included clinical notes, medical records and investigative reports (in texts), CT images (in images) and EHRs (in texts or images). The bioinformatic data dominating forecasting research were genomic data and proteomic data (in sequences).

(2) Data processing

In data preprocessing, the bio-medical data of clinical data (in the format of texts or images) and bioinformatic data (in sequences) were carefully processed into valuable data.

- For *textual clinical data* (e.g., clinical notes, medical records and EHRs), a series of text mining techniques (e.g., data cleaning, tokenization and word stemming) have been used to remove spelling mistakes [198] and irrelevant words [32] and to transform raw data into words or phrases (see model details in Section 3.3.1).
- For *image clinical data* (e.g., CT images and EHRs), corrupted [199], low-contrast [199] and out-of-size images [200] were adjusted or removed to obtain high-quality clinical image data.
- For *bioinformatic data* (e.g., genomic data and proteomic data), data cleaning was conducted to detect and delete duplicate [34] and incomplete sequences [35], leaving valid sequences.

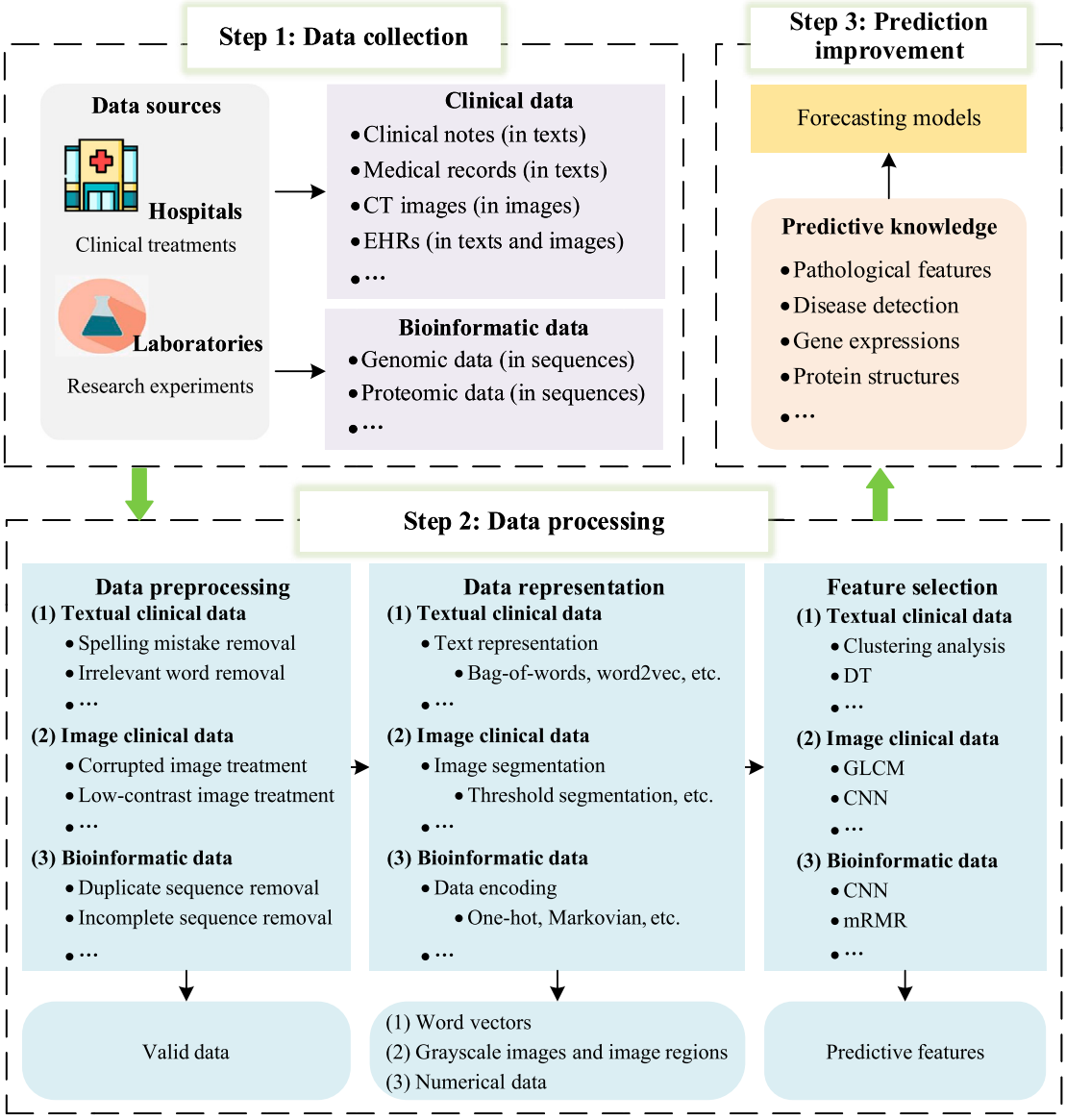

Fig. 18. General framework of bio-medical data-based forecasting research.

In data representation, bio-medical data were transformed into editable formats, with different analysis techniques for different formats.

- For *textual clinical data*, words (or phases) were transformed into word vectors or matrices, via popular text mining technologies of bags-of-words [32] and word2vec [60] (detailed in Section 3.3.1).
- For *image clinical data*, images were transformed into more tractable formats (e.g., grayscale images), cut into image regions and filtered out of meaningless regions, with typical methods of gray processing [201] and threshold segmentation [200] (detailed in Section 3.3.2).
- For *bioinformatic data*, data encoding was performed to transform gene or protein sequences into numerical series, with prevailing encoding methods being one-hot encoding (to encode bioinformatic sequences into 1 or 0) [61] and Markovian encoding (to encode sequences based on the probability distribution) [202].

In feature selection, different analysis methods have been applied to different bio-medical data, to extract the predictive features.

- For *clinical data*, predictive variables were sifted from word vectors (for textual data) via both traditional tools (e.g., clustering analysis [203]; see model details in Section 3.3.1) and AI algorithms (such as DT [204]), and from image regions (for image data) via GLCM [205], CNN [59], etc. (detailed in Section 3.3.2).
- For *bioinformatic data*, existing research has introduced CNNs [61] and minimum redundancy maximum relevance (mRMR, to search the most effective features in terms of having the highest relevance to prediction targets but the lowest redundancy) [206], to extract the hidden predictive knowledge.

(3) Prediction improvement

The predictive features obtained from informative bio-medical data have been introduced as important inputs to improve exist-

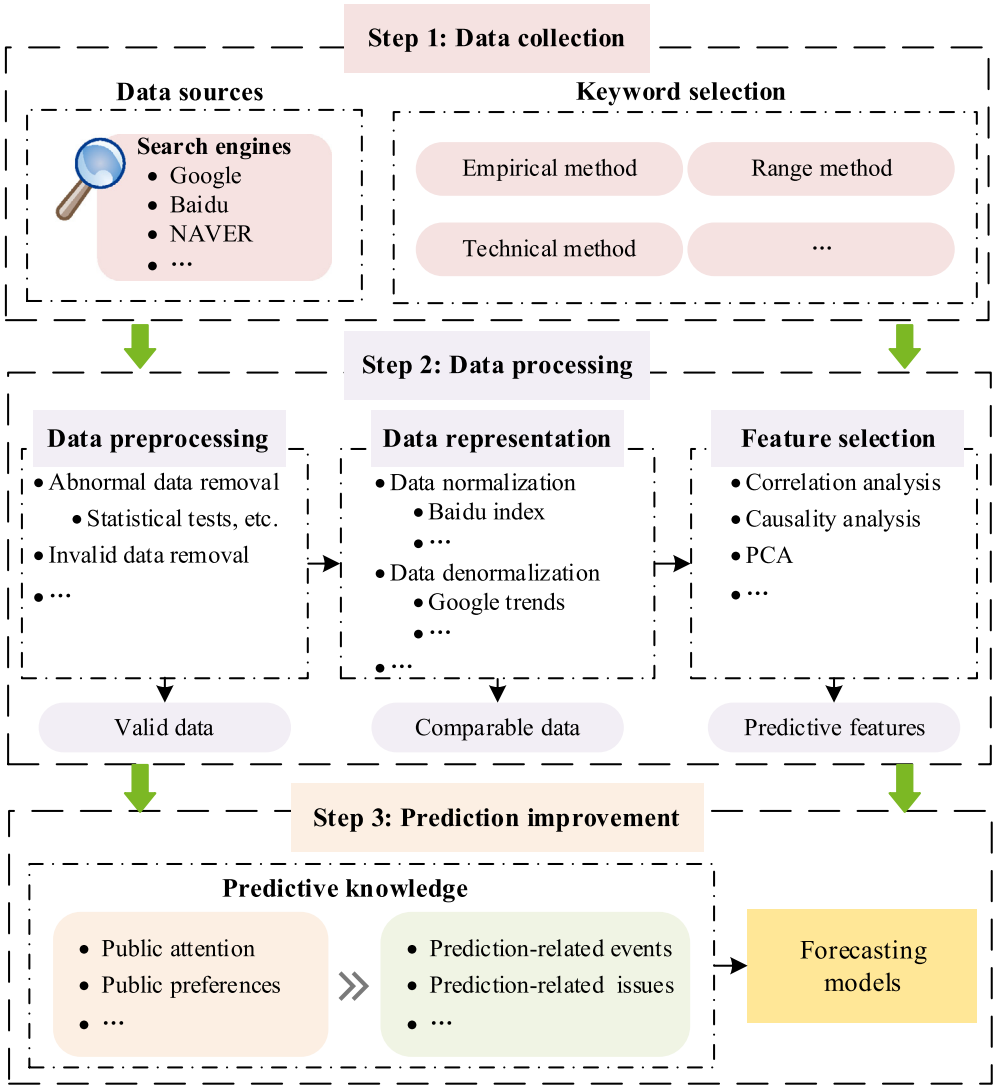

Fig. 19. General framework of web search data-based forecasting research.

ing prediction models, enhancing the prediction accuracy. On the one hand, the clinical data logging disease diagnosis, clinical treatments and patient responses [60], as well as patients' attributes (e.g., age [198], gender [198] and living habits [207]), have helped capture pathological features [67] and facilitated disease detection [60]. On the other hand, the bioinformatic data, detailing lab experiments and the associated results, particularly in gene and protein research, have provided insightful information regarding gene orders [34], gene expressions [61] and protein structures [35], which have been considered to be predictive in the associated predictions.

5.3.2. Web search data

(1) Data collection

First, the search keywords for prediction-related events and/or issues were determined via three main keyword selection methods, i.e., empirical (or experiential) method [6], range (or territorial) method [208] and technical method [209].

- *Empirical method*, a simple but effective method, directly determining prediction target-related keywords based on experts' knowledge and experiences [6].
- *Range method*, an extended version of empirical method, first determining the basic keywords based on empirical method and then adding other related keywords based on the recommendation function of search engines, to get full-scale keywords [208].
- *Technical method*, to filter keywords relevant to prediction targets from a large keyword set, particularly based on correlation analysis [209].

Second, using the determined search keywords, the associated web search data can be obtained from search engines, including Google, Baidu and NAVER.

(2) Data processing

In data preprocessing, to obtain valid web search data, existing forecasting research contrived to detect and remove abnormal samples, for example, using statistical tests (e.g., t test) to find significant deviations from the expected mean [184]; to delete invalid data, particularly those with few search volumes [210].

In data representation, data normalization and data denormalization were performed to transform absolute (e.g., Baidu index) and relative search volumes (e.g., Google trends and NAVER trends) into uniform ranges and original ranges, respectively.

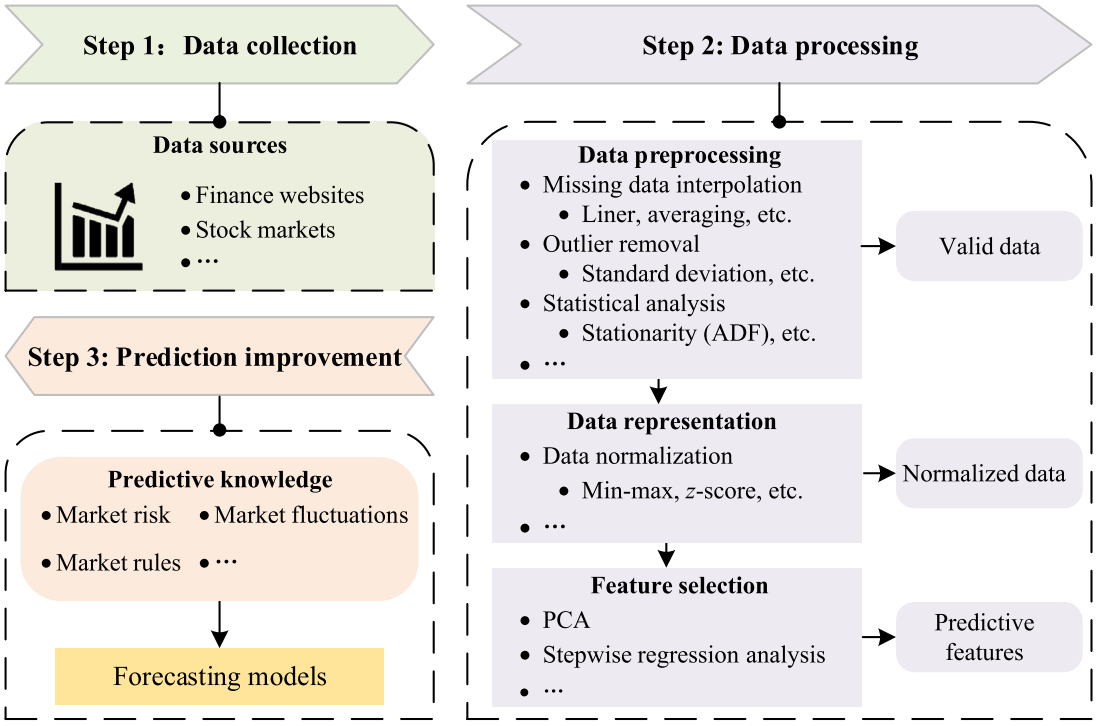

Fig. 20. General framework of stock exchange data-based forecasting research.

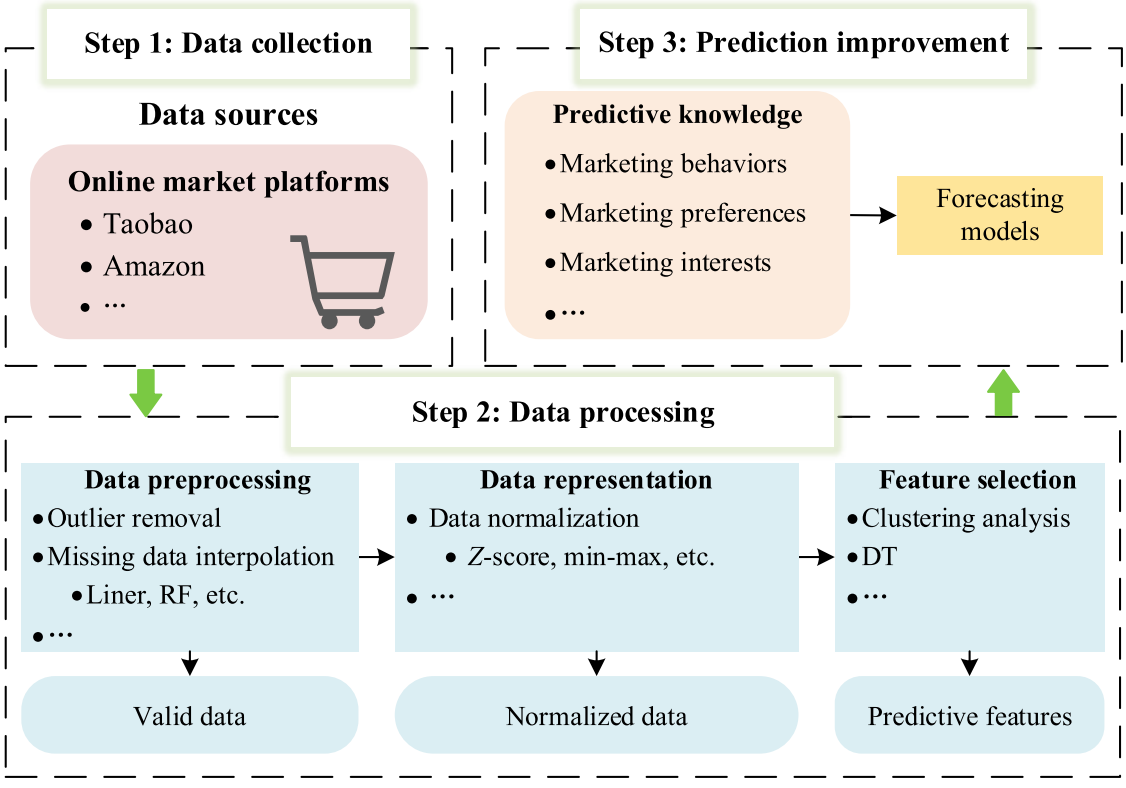

Fig. 21. General framework of online marketing data-based forecasting research.

- *Data normalization*, in which min-max normalization [211] and z-score normalization [212]) were popularly used to normalize the Baidu indexes (absolute search volumes) into uniform ranges for consistency.
- *Data denormalization*, otherwise rescaling the Google trends and NAVER trends (relative search volumes) into their respective original ranges, to reflect the actual difference in magnitude across keywords and time periods [12].

In feature selection, existing research introduced a series of relationship exploration analyses, such as correlation analyses (via Pearson [187], Kendall [12] and Spearman correlation coefficients

[213]) and causality analysis [6], to find predictive web search data in relation to prediction targets. In addition, PCA [214] has also been utilized to obtain the key principal components carrying the majority of the hidden information.

(3) Prediction improvement

Web search data (in terms of search volume for a given keyword) provided insightful new knowledge in prediction, i.e., the public attention [12] and public preferences [210] toward prediction target-related events [208] and/or issues [214], thereby substantially enhancing the prediction accuracy.

5.3.3. Stock exchange data

(1) Data collection

The stock exchange data in prediction, logging the details on every transaction in stock markets (trade time, stock ID, trade price, trade size, buyer, seller, etc.), can be downloaded from finance websites (e.g., the Yahoo finance [185] and Google finance [186]) and stock markets (e.g., the NASDAQ Stock Exchange [73], New York Stock Exchange [73] and Shenzhen Stock Exchange (SZSE) [187]).

(2) Data processing

In data preprocessing, the common operations included missing data interpolation, with popular techniques of linear interpolation [185] and averaging interpolation [215]; outlier identification as those out of standard deviations [216]; data characteristics exploration, e.g., stationarity test using ADF [73].

In data representation, data normalization was commonly conducted to scale stock exchange data into a uniform range to eliminate the effects of data dimensions, with the leading methods of min-max normalization [188] and z-score normalization [187].

In feature selection, existing studies have employed PCA to capture the major information hidden in massive stock exchange data [187] and stepwise regression to determine the predictive features, all of which are revealed to be significantly related to prediction targets but irrelevant to each other [217].

(3) Prediction improvement

The predictive knowledge extracted from stock exchange data ranged from the basic information on individual transactions (regarding stock ID [185], stock trading volume [73], stock price [188], stock index [217], etc.) to insightful knowledge deduced from the information aggregation of massive transactions (reflecting the dynamics of market risk [63], market fluctuations [73], market rules [63], etc.), which have helped better understand the complex system of stock markets and greatly enhance the prediction accuracy.

5.3.4. Online marketing data

(1) Data collection

The online marketing data in prediction, logging the rich information on each online marketing activity (i.e., trade time, price, volume, etc.), were mainly collected from the associated online market platforms (with popular cases of the Taobao [64] and Amazon [218]), through the tools of web crawlers [218] (see method details in Section 3.3).

(2) Data processing

In data preprocessing, existing forecasting studies mainly devoted to outlier removal [219] and missing data interpolation, based on not only traditional techniques (e.g., linear interpolation method [220]) but also machine learning algorithms (e.g., RF, to search the optimal interpolations for missing data by training a series of randomly generated trees [219]).

In data representation, data normalization was commonly performed to remove data dimensions, via z-score normalization [221], min-max normalization [222], etc.

In feature selection, not only statistical models (e.g., clustering analysis to group features [64]) but also AIs (e.g., DT to adaptively search informative features based on Gini impurity [200]) have been employed to extract the predictive features from massive online marketing data.

(3) Prediction improvement

Based on detailed log information on online marketing (regarding consumers [219], trade time [64], products [218]), trade prices [218], trade volumes [64], etc.), new knowledge can be deduced regarding customers' marketing behaviors, preferences and interests [64], which appeared to be powerful predictors in the related predictions.

5.3.5. Other log data

In addition to the above data, various other log data have also been applied to forecasting research, such as web log data, smart card data and highway traffic data. To withdraw the new knowledge from each of them, the three steps of data collection, data processing and prediction improvement were conducted as follows.

(1) Web log data

In *data collection*, any activities on a web would be automatically recorded and collected by the associated web servers, forming web log data. In *data processing*, data preprocessing was performed to filter out unrelated log records [190] and fill null values via probability smoothing algorithm [223], etc.; data representation was conducted to provide descriptive statistics (e.g., frequency, volume and distribution) for each activity type on the web, e.g., web access [190], visit/browse [62], click [224] and attack [225], and then process them using data normalization (e.g., z-score normalization [226]) and index construction [224]; feature selection has introduced not only statistical methods (i.e., correlation analysis and Granger causality analysis [86]) but also machine learning algorithms (e.g., DT [190]) to select predictive features from web log statistics. In *prediction improvement*, the predictive features were examined to be access frequency [190], visit volume [62], visit preference [86] and click distribution [224] and the event or topic of the associated web [225], which were widely used as important inputs in prediction models to reflect user interests [190] and usage patterns [109].

(2) Smart card data

In *data collection*, the smart card data in prediction were generated upon the usage of smart cards (e.g., bus smart cards in daily traffic and credit cards in transactions) [191]. In *data processing*, data cleaning was taken to remove missing data, leaving valid data; data representation focused on data normalization, particularly via min-max normalization [227]; feature selection used PCA to extract the main information [191]. In *prediction improvement*, the predictive features, extracted from the log details on timestamps [227], card ID [227], commodity and trade price and quantity [191], etc., effectively reflected consumer interests and patterns (e.g., spatiotemporal distribution) [227] and improved forecasting models.

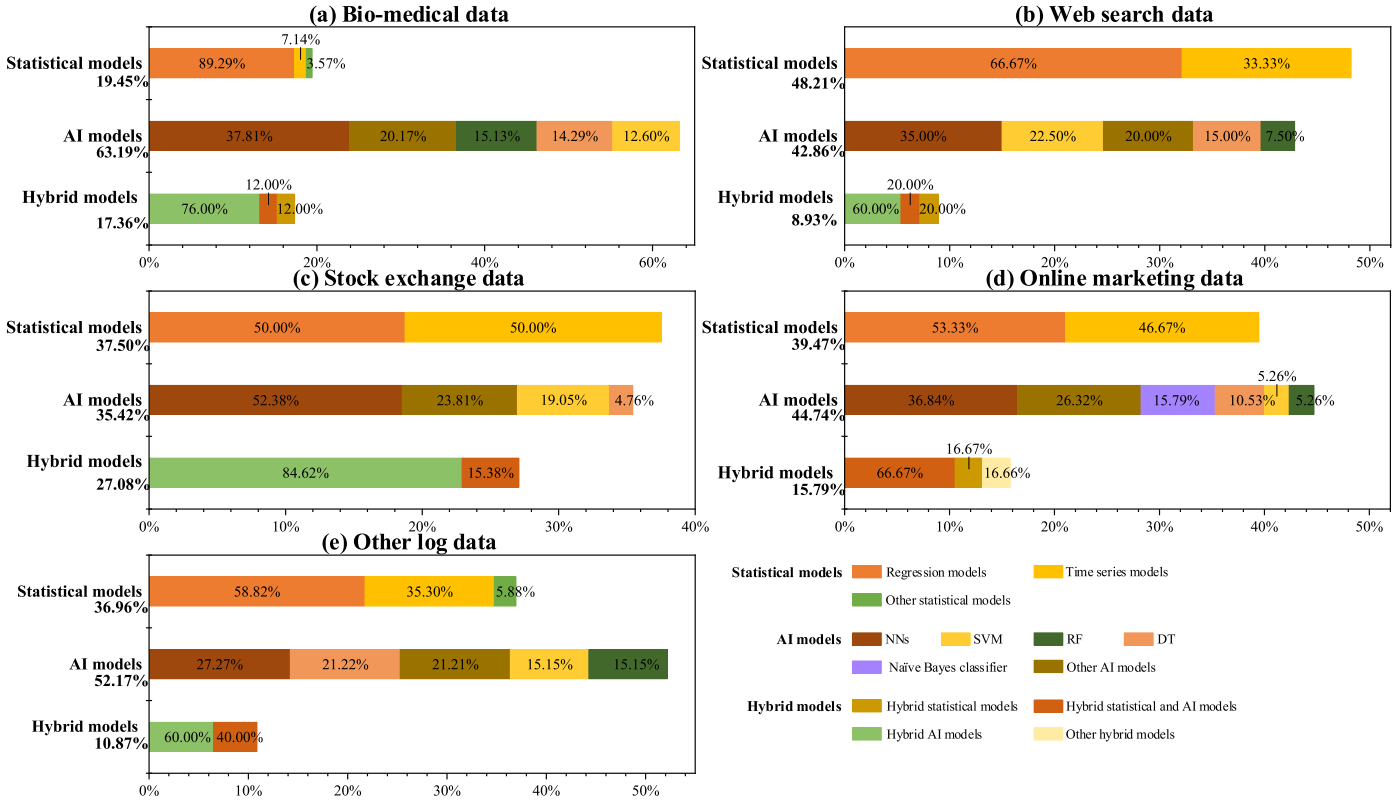

Fig. 22. Distributions of forecasting models in forecasting research using log data, i.e., bio-medical data (**a**), web search data (**b**), stock exchange data (**c**), online marketing data (**d**) and other log data (**e**).

(3) Highway traffic data

In *data collection*, the highway traffic data in forecasting research mainly focused on highway traffic transactions (e.g., highway toll payments) [74]. In *data processing*, data cleaning was commonly operated to remove missing data [74]; clustering analysis was popularly employed to group valid highway traffic data into valuable features [74]; promising machine learning methods (e.g., DT) were also introduced to adaptively search effective features [192]. In *prediction improvement*, based on the basic information on highway traffic transactions (e.g., vehicle information, toll amount, and time) [192], the extracted predictive features finely revealed the trajectory dynamics and spatiotemporal distributions of driving behaviors [192], thereby greatly promoting the related predictions.

5.4. Forecasting models

As shown in Fig. 22, AI models dominated log data-based forecasting research (representing 51.53% of the associated articles), followed by statistical models and hybrid models (32.52% and 15.95%, respectively).

Among statistical models, regression analysis had a wide application in prediction (representing 66.35% of log data-based articles using statistical models), particularly LR for web search data [184] and stock exchange data [73], and logistic regression for bio-medical data [32] and other log data [191]. In addition, time series models also made a large contribution to log data-based prediction (31.73%), with the typical examples of seasonal autoregressive integrated moving average (SARIMA) for web search data [228] and bio-medical data [228], vector autoregression (VAR) for stock exchange data [229], and ARIMA for online marketing data [218].

AI models prevailed in log data-based forecasting, in which the dominant models for all the types of log data included NNs (accounting for 37.07% of log data-based articles using AIs) [34], SVM (14.66%) [153] and DT (14.22%) [6]. Moreover, some other AI models have also been shown to be superior in modeling a particular type of log data, such as RF for bio-medical data (representing 15.13% of bio-medical data-based studies using AIs) [230] and web search data (7.50%) [231], DT for stock exchange data (representing 4.76% of stock exchange data-based studies using AIs) [232] and naive Bayes for online marketing data (representing 15.79% of online marketing data-based studies using AIs) [233].

Hybrid models, a rising star in forecasting research, have recently been introduced into log data-based prediction. The popular types were majorly formulated based on AI models (representing 66.67% of log data-based articles using hybrid models) [188], statistical and AI models (22.22%) [234] and statistical models (9.26%) [111].

5.5. Major findings and future directions

With the updates of data processing and storage technologies, activities or operations can be recorded in detail in terms of log data and have promoted forecasting research. However, there is still much room to improve log data-based forecasting studies. For data types, some valuable big data have few applications in forecasting research, for example, the case-specific, large-scale logs on operations of prevention, infection, treatment, control, etc. of major epidemic diseases (such as the SARS outbreak in 2013 and COVID-19 outbreak in 2019). The hidden reason might be that these log details are mainly in the control of official institutions (e.g., medical and health organizations and the government), preventing data sharing due to privacy concerns [235]. Therefore, a reciprocal cooperation between the government and academia could be a promising way to address data limitations and take full use of detailed public health logs, thereby enhancing the prediction accuracy [236].

For forecasting hotspots, log data have made great contributions to social prediction (representing 49.00% of log data-based articles, particularly for market activities) and biology prediction (46.03%, particularly for biomedicine). However, such a valuable type of big data has far fewer applications in the domain of nature (4.97%) and has rarely served some forecasting hotspots, e.g., weather factors and material properties. The hidden reason might be that the log data in prediction mainly was concerned with human activities (related to social prediction) and medical activities (biological prediction), while the activities related to natural prediction (particularly lab experiments in natural science research) are strongly suggested to be recorded in detail and carefully analyzed to promote natural prediction.

For analysis technologies, both statistical analyses and AI techniques have been introduced to process log data and extract the hidden predictive knowledge. However, some other even more popular and effective log data processing techniques had few applications in log data-based prediction. For example, data desensitization techniques have been extensively used in the research of financial market supervision [237], telecom customer analysis [238], distributed energy system management [239], etc. to obfuscate sensitive information for addressing privacy concerns and data limitations, which can also be introduced to promote forecasting research [240].

For forecasting models, the most popular models in existing log data-based forecasting research were AI models (accounting for 51.53% of log data-based articles), while the application of hybrid models in prediction was still at an early stage in terms of a small number of publications (15.95%). However, such promising models have shown significant power in modeling complicated data (particularly big data) [234]. Therefore, there is a large opportunity to forecast research improvements by formulating and employing effective variants of hybrid models according to the unique features of log data.

6. Conclusion

With the boom in Internet techniques and computer science, a variety of big data have been introduced into forecasting research to bring insightful knowledge and improve forecasting models. Generally, big data-based forecasting research has a relatively long history (since 2004, just three years after the dawn of the big data era) and has experienced rapid growth, particularly since 2014. The major publication sources for our study were journal papers (covering a total of 1,913 journals and accounting for 79.48% of the total articles). The top influential contributors included the USA (representing 29.19% of the total articles) and China (28.57%), and the strongest international cooperative relationship lied between the USA and China (representing 3.93% of the total cooperative articles).

This paper is the first attempt to review full-scale types of big data used in forecasting research. In particular, for each type of big data, an overall review is provided detailing the specific types and sources (i.e., what data to use), forecasting hotspots (where the data dominated) and analysis and forecasting models (how to use the data to improve prediction), as well as the associated major findings and future directions.

For *what* big data to use in prediction, the big data in forecasting research mainly fall into three categories: (1) user-generated content data (representing for 13.94% of the total articles, generated from the users on social media or other web platforms), such as online textual data and online photo data, which brought new knowledge into prediction regarding users' opinions and attention to prediction target-related issues or events; (2) device-monitored data (54.14%, monitored by devices), including meteorological data, smart meter data, traffic flow data, etc. which provided the associated sensor-level, real-time dynamics, thereby substantially enhancing the spatiotemporal resolution of forecasting models; (3) log data (31.92%, recording operations or activities in detail), involving bio-medical data, web search data, stock exchange data, online marketing data, etc., from which general rules and insightful knowledge can be deduced and used to improve prediction.

For *where* big data dominated, the forecasting hotspots included: social prediction for human behaviors, market factors, social events, transportation, etc. (jointly accounting for 49.70% of the total articles); natural prediction for weather factors, environmental factors, engineering issues, material properties, etc. (29.87%); biological prediction for biomedicine, biotechnology, animal and plant science, etc. (20.43%). Notably, because they carry different knowledge, different types of big data have different forecasting hotspots: UGC data, shared by individual netizens and reflecting public emotions, opinions and attention, dominated social prediction (representing 88.23% of UGC data-based articles), particularly for social events and human behaviors; device data, mainly monitoring engineering issues and weather environments in the nature domain, prevailed in natural prediction (representing 51.90% of device data-based articles); log data, majorly recording human activities and marketing behaviors, focused on social prediction (representing 49.00% of log data-based articles).

For *how* to use big data, three major steps are taken, i.e., data collection (from different sources), data processing (to extract predictive knowledge) and prediction improvement (using big data). In data processing, a set of analysis technologies have been employed to extract the insightful information hidden in big data through the following three sub-steps: data preprocessing, to identify and treat unrelated, duplicated, missing and abnormal values; data representation, to transform big data into a structured format (with typical operations of text representation for textual data, image segmentation for photo data and data encoding for sequence data) and into a comparable unit (using data normalization or denormalization); feature selection, to extract useful information through word vector selection for textual data (with popular tools of LDA and sentimental analysis), image feature selection for photo data (with color histogram and GLCM), informative feature extraction from massive features (with PCA and clustering analysis); and relationship exploration to find prediction target-related features (with correlation analysis and Granger causality analysis). In prediction improvement, the predictive knowledge extracted in data processing is put into forecasting models as important inputs. The prevailing forecasting models included AIs (accounting for 60.94% of the total articles), with popular cases of NNs (e.g., back propagation neural networks (BPNN), CNN and extreme learning machine (ELM)), SVM (e.g., SVR and least square SVR (LSSVR)), DT and RF; statistical models (23.21%), including regression analyses (e.g., LR and logistic regression) and time series models (e.g., ARIMA, SARIMA and VAR); hybrid models (15.85%), a rising star in prediction research, particularly those combining AI models and statistical and AI models (with a promising case of decomposition and ensemble models).

Even with an encouraging research improvement, there still exists ample room to develop big data-based forecasting research by enriching data types, extending prediction applications and improving data analysis technologies and forecasting models. For data types, some valuable big data can also be widely introduced to enrich big data-based forecasting research, such as the UGC data of audio and video data, device data of Wi-Fi data and Bluetooth data, and case-specific, large-scale log data for major epidemic diseases (particularly COVID-19). For prediction tasks, UGC data have few applications in the natural prediction of engineering issues and material properties, device data have been used far less frequently in biological prediction (particularly for biotechnology), and log data have paid little attention to the natural prediction of weather

factors and material properties, thereby implying a large opportunity for prediction improvement. For data analyses, some emerging but powerful analysis technologies that have already been used in other research areas are strongly recommended. Promising examples are latent aspect rating analysis and semantic-based image retrieval to extract new information from online textual and image data, respectively, data compression technologies to accelerate the transmission and processing of device data, and data desensitization to address privacy concerns in log data. For forecasting models, due to the nonstationary, nonlinearity and complexity of big data, hybrid models—which combine a series of models to address the shortcomings of another model by taking advantage of the other models specifically in modeling a complex system—are strongly recommended, particularly in online photo data-based prediction (where hybrid models have not yet been employed).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by grants from the National Natural Science Foundation of China (NSFC Nos. 72004144; 71971007).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.bdr.2021.100289.

References

[1] A. Ali, J. Qadir, U.R. Rasool, A. Sathiaseelan, A. Zwitter, J. Crowcroft, Big data for development: applications and techniques, Big Data Anal. 1 (1) (2016) 1–24, https://doi.org/10.1186/s41044-016-0002-4.
[2] D. Laney, 3D data management: controlling data volume, velocity and variety, META Gr. Res. Note. 6 (70) (2001) 1.
[3] X. Jin, B.W. Wah, X. Cheng, Y. Wang, Significance and challenges of big data research, Big Data Res. 2 (2) (2015) 59–64, https://doi.org/10.1016/j.bdr.2015.01.006.
[4] N. Elgendy, A. Elragal, Big data analytics: a literature review paper, in: Industrial Conference on Data Mining, Springer, 2014, pp. 214–227.
[5] I.A.T. Hashem, I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani, S. Ullah Khan, The rise of "big data" on cloud computing: review and open research issues, Inf. Syst. 47 (2015) 98–115, https://doi.org/10.1016/j.is.2014.07.006.
[6] L. Yu, Y. Zhao, L. Tang, Z. Yang, Online big data-driven oil consumption forecasting with Google trends, Int. J. Forecast. 35 (1) (2019) 213–223, https://doi.org/10.1016/j.ijforecast.2017.11.005.
[7] R. Elshawi, S. Sakr, D. Talia, P. Trunfio, Big data systems meet machine learning challenges: towards big data science as a service, Big Data Res. 14 (2018) 1–11, https://doi.org/10.1016/j.bdr.2018.04.004.
[8] O. Schaer, N. Kourentzes, R. Fildes, Demand forecasting with user-generated online information, Int. J. Forecast. 35 (1) (2019) 197–212, https://doi.org/10.1016/j.ijforecast.2018.03.005.
[9] M.M. Gobble, Big data: the next big thing in innovation, Res. Manag. 56 (2013) 64–67, https://doi.org/10.5437/08956308X5601005.
[10] M. Nilashi, A. Ahani, M.D. Esfahani, E. Yadegaridehkordi, S. Samad, O. Ibrahim, N.M. Sharef, E. Akbari, Preference learning for eco-friendly hotels recommendation: a multi-criteria collaborative filtering approach, J. Clean. Prod. 215 (2019) 767–783, https://doi.org/10.1016/j.jclepro.2019.01.012.
[11] J. Bao, P. Liu, S.V. Ukkusuri, A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data, Accid. Anal. Prev. 122 (2019) 239–254, https://doi.org/10.1016/j.aap.2018.10.015.
[12] A.F.W. Ho, B.Z.Y.S. To, J.M. Koh, K.H. Cheong, Forecasting hospital emergency department patient volume using Internet search data, IEEE Access 7 (2019) 93387–93395, https://doi.org/10.1109/ACCESS.2019.2928122.
[13] S. Seo, D. Kang, Study on predicting sentiment from images using categorical and sentimental keyword-based image retrieval, J. Supercomput. 72 (9) (2016) 3478–3488, https://doi.org/10.1007/s11227-015-1510-0.
[14] T. Ahmad, H. Chen, W.A. Shah, Effective bulk energy consumption control and management for power utilities using artificial intelligence techniques under conventional and renewable energy resources, Int. J. Electr. Power Energy Syst. 109 (2019) 242–258, https://doi.org/10.1016/j.ijepes.2019.02.023.
[15] K. Li, W. Lu, C. Liang, B. Wang, Intelligence in tourism management: a hybrid FOA-BP method on daily tourism demand forecasting with web search data, Mathematics 7 (6) (2019) 531, https://doi.org/10.3390/MATH7060531.
[16] M.M. Tadesse, H. Lin, B. Xu, L. Yang, Personality predictions based on user behavior on the Facebook social media platform, IEEE Access 6 (2018) 61959–61969, https://doi.org/10.1109/ACCESS.2018.2876502.
[17] V.V. Nhlabano, P.E.N. Lutu, Impact of text pre-processing on the performance of sentiment analysis models for social media data, in: 2018 International Conference on Advances in Big Data, 2018, pp. 1–6.
[18] A.Y.L. Chong, E. Ch'ng, M.J. Liu, B. Li, Predicting consumer product demands via big data: the roles of online promotional marketing and online reviews, Int. J. Prod. Res. 55 (17) (2017) 5142–5156, https://doi.org/10.1080/00207543.2015.1066519.
[19] M.J. Schneider, S. Gupta, Forecasting sales of new and existing products using consumer reviews: a random projections approach, Int. J. Forecast. 32 (2) (2016) 243–256, https://doi.org/10.1016/j.ijforecast.2015.08.005.
[20] L. Mohan, S. Elayidom, Predicting the winner of Delhi assembly election, 2015 from sentiment analysis on Twitter data-a bigdata perspective, Int. Arab J. Inf. Technol. 16 (5) (2019) 833–842.
[21] Y.G. Petalas, A. Ammari, P. Georgakis, C. Nwagboso, A big data architecture for traffic forecasting using multi-source information, in: International Workshop of Algorithmic Aspects of Cloud Computing, Springer, 2016, pp. 65–83.
[22] P. Jiang, L. Liu, L. Cui, H. Li, Y. Shi, Congestion prediction of urban traffic employing SRBDP, in: 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications, IEEE, 2017, pp. 1099–1106.
[23] A.J. Hussain, P. Liatsis, M. Khalaf, H. Tawfik, H. Al-Asker, A dynamic neuralnetwork architecture with immunology inspired optimization for weather data forecasting, Big Data Res. 14 (2018) 81–92, https://doi.org/10.1016/j.bdr.2018.04.002.
[24] S.N. Qasem, S. Samadianfard, H.S. Nahand, A. Mosavi, S. Shamshirband, K.W. Chau, Estimating daily dew point temperature using machine learning algorithms, Water 11 (3) (2019) 582, https://doi.org/10.3390/w11030582.
[25] W. Yuan, K. Wang, X. Bo, L. Tang, J. Wu, A novel multi-factor & multi-scale method for $PM_{2.5}$ concentration forecasting, Environ. Pollut. 255 (2019) 113187, https://doi.org/10.1016/j.envpol.2019.113187.
[26] J. Lee, J.H. Lee, Constructing efficient regional hazardous weather prediction models through big data analysis, J. Intell. Fuzzy Syst. 16 (1) (2016) 1–12, https://doi.org/10.5391/ijfis.2016.16.1.1.
[27] L. Ren, J. Cui, Y. Sun, X. Cheng, Multi-bearing remaining useful life collaborative prediction: a deep learning approach, J. Manuf. Syst. 43 (2017) 248–256, https://doi.org/10.1016/j.jmsy.2017.02.013.
[28] S. Baek, D.Y. Kim, Abrupt variance and discernibility analyses of multi-sensor signals for fault pattern extraction, Comput. Ind. Eng. 128 (2019) 999–1007, https://doi.org/10.1016/j.cie.2018.06.019.
[29] L. Kang, H.L. Du, H. Zhang, W.L. Ma, Systematic research on the application of steel slag resources under the background of big data, Complexity 2018 (2018) 6703908, https://doi.org/10.1155/2018/6703908.
[30] S. Guo, J. Yu, X. Liu, C. Wang, Q. Jiang, A predicting model for properties of steel using the industrial big data based on machine learning, Comput. Mater. Sci. 160 (2019) 95–104, https://doi.org/10.1016/j.commatsci.2018.12.056.
[31] X. Wang, S. Yang, Y. Zhao, Y. Wang, Improved pore structure prediction based on MICP with a data mining and machine learning system approach in Mesozoic strata of Gaoqing field, Jiyang depression, J. Pet. Sci. Eng. 171 (2018) 362–393, https://doi.org/10.1016/j.petrol.2018.07.057.
[32] I. Segura-Bedmar, C. Colón-Ruíz, M.Á. Tejedor-Alonso, M. Moro-Moro, Predicting of anaphylaxis in big data EMR by exploring machine learning approaches, J. Biomed. Inform. 87 (2018) 50–59, https://doi.org/10.1016/j.jbi.2018.09.012.
[33] S. Chae, S. Kwon, D. Lee, Predicting infectious disease using deep learning and big data, Int. J. Environ. Res. Public Health 15 (8) (2018) 1596, https://doi.org/10.3390/ijerph15081596.
[34] J.W. Chang, Y. Ding, M. Tahir ul Qamar, Y. Shen, J. Gao, L.L. Chen, A deep learning model based on sparse auto-encoder for prioritizing cancer-related genes and drug target combinations, Carcinogenesis 40 (5) (2019) 624–632, https://doi.org/10.1093/carcin/bgz044.
[35] J.Y. An, Z.H. You, Y. Zhou, D.F. Wang, Sequence-based prediction of protein-protein interactions using gray wolf optimizer-based relevance vector machine, Evol. Bioinform. 15 (2019) 1–10, https://doi.org/10.1177/1176934319844522.
[36] M.O.D. Rizwan, R.J.R. Raj, M. Vasudev, A novel approach for time series data forecasting based on ARIMA model for marine fishes, in: 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies, IEEE, 2017, pp. 1–4.
[37] K.E. Anderson, N.F. Glenn, L.P. Spaete, D.J. Shinneman, D.S. Pilliod, R.S. Arkle, S.K. McIlroy, D.W.R. Derryberry, Estimating vegetation biomass and cover across large plots in shrub and grass dominated drylands using terrestrial li-

dar and machine learning, Ecol. Indic. 84 (2018) 793–802, https://doi.org/10.1016/j.ecolind.2017.09.034.
[38] H. Hassani, E.S. Silva, Forecasting with big data: a review, Ann. Data Sci. 2 (1) (2015) 5–19, https://doi.org/10.1007/s40745-015-0029-9.
[39] K. Grolinger, M.A.M. Capretz, L. Seewald, Energy consumption prediction with big data: balancing prediction accuracy and computational resources, in: 2016 IEEE International Congress on Big Data (BigData Congress), 2016, pp. 157–164.
[40] Y. Wang, Q. Chen, T. Hong, C. Kang, Review of smart meter data analytics: applications, methodologies, and challenges, IEEE Trans. Smart Grid 10 (2019) 3125–3148, https://doi.org/10.1109/TSG.2018.2818167.
[41] S.J. Chua, S. Wrigley, C. Hair, R. Sahathevan, Prediction of delirium using data mining: a systematic review, J. Clin. Neurosci. 91 (2021) 288–298, https://doi.org/10.1016/j.jocn.2021.07.029.
[42] A. Sammani, A.F. Baas, F.W. Asselbergs, A.S.J.M. Te Riele, Diagnosis and risk prediction of dilated cardiomyopathy in the era of big data and genomics, J. Clin. Med. 10 (2021) 921, https://doi.org/10.3390/jcm10050921.
[43] S. Yang, L.G. Stansbury, P. Rock, T. Scalea, P.F. Hu, Linking big data and prediction strategies: tools, pitfalls, and lessons learned, Crit. Care Med. 47 (2019) 840–848, https://doi.org/10.1097/CCM.0000000000003739.
[44] D. Pavlyuk, Feature selection and extraction in spatiotemporal traffic forecasting: a systematic literature review, Eur. Transp. Res. Rev. 11 (2019), https://doi.org/10.1186/s12544-019-0345-9.
[45] T. Hong, S. Fan, Probabilistic electric load forecasting: a tutorial review, Int. J. Forecast. 32 (3) (2016) 914–938, https://doi.org/10.1016/j.ijforecast.2015.11.011.
[46] M.A. Al-Garadi, M.R. Hussain, N. Khan, G. Murtaza, H.F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H.A. Khattak, A. Gani, Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges, IEEE Access 7 (2019) 70701–70718, https://doi.org/10.1109/ACCESS.2019.2918354.
[47] I.R.S. Agostino, W.V. da Silva, C. Pereira da Veiga, A.M. Souza, Forecasting models in the manufacturing processes and operations management: systematic literature review, J. Forecast. 39 (2020) 1043–1056, https://doi.org/10.1002/for.2674.
[48] X. Ren, X. Li, K. Ren, J. Song, Z. Xu, K. Deng, X. Wang, Deep learning-based weather prediction: a survey, Big Data Res. 23 (2021) 100178, https://doi.org/10.1016/j.bdr.2020.100178.
[49] M. Fathi, M.H. Kashani, S.M. Jameii, E. Mahdipour, Big data analytics in weather forecasting: a systematic review, Arch. Comput. Methods Eng. (2021) 1–29, https://doi.org/10.1007/s11831-021-09616-4.
[50] J.T. Senders, P.C. Staples, A.V. Karhade, M.M. Zaki, W.B. Gormley, M.L.D. Broekman, T.R. Smith, O. Arnaout, Machine learning and neurosurgical outcome prediction: a systematic review, World Neurosurg. 109 (2018) 476–486, https://doi.org/10.1016/j.wneu.2017.09.149.
[51] M. Kaur, H. Gulati, H. Kundra, Data mining in agriculture on crop price prediction: techniques and applications, Int. J. Comput. Appl. 99 (2014) 1–3, https://doi.org/10.5120/17422-8273.
[52] S. Saran, P. Chaudhary, A. Uttam, S. Gupta, Analysis and optimization of groundwater distribution using SVM and neural networks, in: Proceedings of the International Conference on Innovative Computing & Communications, 2020.
[53] R. Yang, L. Yu, Y. Zhao, H. Yu, G. Xu, Y. Wu, Z. Liu, Big data analytics for financial market volatility forecast based on support vector machine, Int. J. Inf. Manag. 50 (2020) 452–462.
[54] H. Es-Samaali, A. Outchakoucht, J.P. Leroy, A blockchain-based access control for big data, Int. J. Comput. Networks Commun. Secur. 5 (7) (2017) 137–147.
[55] I. Rahimi, R. Behmanesh, A. Ahmadi, Scientometric analysis of scheduling in renewable energy: a keyword and citation analysis, J. Energy Power Technol. 1 (2019) 1–15, https://doi.org/10.21926/jept.1904004.
[56] I. Yaqoob, I.A.T. Hashem, A. Gani, S. Mokhtar, E. Ahmed, N.B. Anuar, A.V. Vasilakos, Big data: from beginning to future, Int. J. Inf. Manag. 36 (2016) 1231–1247, https://doi.org/10.1016/j.ijinfomgt.2016.07.009.
[57] J. Li, Z. Xu, H. Xu, L. Tang, L. Yu, Forecasting oil price trends with sentiment of online news articles, Asia-Pac. J. Oper. Res. 34 (2) (2017) 1740019, https://doi.org/10.1142/S021759591740019X.
[58] X.Y. Ni, H. Huang, W.P. Du, Relevance analysis and short-term prediction of $PM_{2.5}$ concentrations in Beijing based on multi-source data, Atmos. Environ. 150 (2017) 146–161, https://doi.org/10.1016/j.atmosenv.2016.11.054.
[59] I. Lana, J. Del Ser, I.I. Olabarrieta, Understanding daily mobility patterns in urban road networks using traffic flow analytics, in: NOMS 2016-2016 IEEE/IFIP Network Operations and Management Symposium, IEEE, 2016, pp. 1157–1162.
[60] Y. Hao, M. Usama, J. Yang, M.S. Hossain, A. Ghoneim, Recurrent convolutional neural network based multimodal disease risk prediction, Future Gener. Comput. Syst. 92 (2019) 76–83, https://doi.org/10.1016/j.future.2018.09.031.
[61] L. Xue, B. Tang, W. Chen, J. Luo, Prediction of CRISPR sgRNA activity using a deep convolutional neural network, J. Chem. Inf. Model. 59 (1) (2019) 615–624, https://doi.org/10.1021/acs.jcim.8b00368.
[62] Z. Xie, G. Liu, J. Wu, Y. Tan, Big data would not lie: prediction of the 2016 Taiwan election via online heterogeneous information, EPJ Data Sci. 7 (1) (2018) 32, https://doi.org/10.1140/epjds/s13688-018-0163-7.
[63] J. Qi, L. Yi, Y. Chen, Forecasting market risk using ultra-high-frequency data and scaling laws, Quant. Finance 18 (12) (2018) 2085–2099, https://doi.org/10.1080/14697688.2018.1453166.
[64] J. Zhang, A. Simeone, P. Gu, B. Hong, Product features characterization and customers' preferences prediction based on purchasing data, CIRP Ann. 67 (1) (2018) 149–152, https://doi.org/10.1016/j.cirp.2018.04.020.
[65] A. Nigam, H.K. Dambanemuya, M. Joshi, N.V. Chawla, Harvesting social signals to inform peace processes implementation and monitoring, Big Data 5 (4) (2017) 337–355, https://doi.org/10.1089/big.2017.0055.
[66] S. Liu, Y. Ji, D. Zhang, Y. Yuan, J. Gong, R. Wang, An online prediction algorithm of traffic in big data based on the storm, in: 2017 Fifth International Conference on Advanced Cloud and Big Data, IEEE, 2017, pp. 129–134.
[67] C. Yao, S. Wu, Z. Liu, P. Li, A deep learning model for predicting chemical composition of gallstones with big data in medical Internet of Things, Future Gener. Comput. Syst. 94 (2019) 140–147, https://doi.org/10.1016/j.future.2018.11.011.
[68] A. Lenhart, S. Fox, Twitter and Status Updating, Pew Internet & American Life Project, Washington, DC, 2009.
[69] S. Mohan, P. Saranya, A novel bagging ensemble approach for predicting summertime ground-level ozone concentration, J. Air Waste Manage. Assoc. 69 (2) (2019) 220–233, https://doi.org/10.1080/10962247.2018.1534701.
[70] K. Nam, N. Seong, Financial news-based stock movement prediction using causality analysis of influence in the Korean stock market, Decis. Support Syst. 117 (2019) 100–112, https://doi.org/10.1016/j.dss.2018.11.004.
[71] T. Cerquitelli, G. Malnati, D. Apiletti, Exploiting scalable machine-learning distributed frameworks to forecast power consumption of buildings, Energies 12 (15) (2019) 2933, https://doi.org/10.3390/en12152933.
[72] X. Lyu, C. Jiang, Y. Ding, Z. Wang, Y. Liu, Sales prediction by integrating the heat and sentiments of product dimensions, Sustain. 11 (3) (2019) 913, https://doi.org/10.3390/su11030913.
[73] U. Khan, F. Aadil, M.A. Ghazanfar, S. Khan, N. Metawa, K. Muhammad, I. Mehmood, Y. Nam, A robust regression-based stock exchange forecasting and determination of correlation between stock markets, Sustain. 10 (10) (2018) 3702, https://doi.org/10.3390/su10103702.
[74] S. hun Park, S. min Kim, Y. guk Ha, Highway traffic accident prediction using VDS big data analysis, J. Supercomput. 72 (7) (2016) 2815–2831, https://doi.org/10.1007/s11227-016-1624-z.
[75] B. Li, K.C.C. Chan, C. Ou, S. Ruifeng, Discovering public sentiment in social media for predicting stock movement of publicly listed companies, Inf. Syst. 69 (2017) 81–92, https://doi.org/10.1016/j.is.2016.10.001.
[76] J. Luo, M. Wu, D. Gopukumar, Y. Zhao, Big data application in biomedical research and health care: a literature review, Biomed. Inform. Insights 8 (2016) 1–10, https://doi.org/10.4137/BII.S31559.
[77] M. Chen, S. Mao, Y. Liu, Big data: a survey, Mob. Netw. Appl. 19 (2014) 171–209, https://doi.org/10.1007/s11036-013-0489-0.
[78] M. Salehan, D.J. Kim, Predicting the performance of online consumer reviews: a sentiment mining approach to big data analytics, Decis. Support Syst. 81 (2016) 30–40, https://doi.org/10.1016/j.dss.2015.10.006.
[79] J. Devaraj, R. Madurai Elavarasan, G.M. Shafiullah, T. Jamal, I. Khan, A holistic review on energy forecasting using big data and deep learning models, Int. J. Energy Res. 45 (9) (2021) 13489–13530, https://doi.org/10.1002/er.6679.
[80] M. Akhtar, S. Moridpour, A review of traffic congestion prediction using artificial intelligence, J. Adv. Transp. 2021 (2021) 1–18, https://doi.org/10.1155/2021/8878011.
[81] A. Zendehboudi, M.A. Baseer, R. Saidur, Application of support vector machine models for forecasting solar and wind energy resources: a review, J. Clean. Prod. 199 (2018) 272–285, https://doi.org/10.1016/j.jclepro.2018.07.164.
[82] J.B. Kristensen, T. Albrechtsen, E. Dahl-Nielsen, M. Jensen, M. Skovrind, T. Bornakke, Parsimonious data: how a single Facebook like predicts voting behavior in multiparty systems, PLoS ONE 12 (9) (2017) e0184562, https://doi.org/10.1371/journal.pone.0184562.
[83] X. Ren, X. Chen, Discovery and dynamic prediction of user's interest based on ARIMA, in: 2017 Portland International Conference on Management of Engineering and Technology, IEEE, 2017, pp. 1–8.
[84] Z.P. Fan, Y.J. Che, Z.Y. Chen, Product sales forecasting using online reviews and historical sales data: a method combining the Bass model and sentiment analysis, J. Bus. Res. 74 (2017) 90–100, https://doi.org/10.1016/j.jbusres.2017.01.010.
[85] S. Tariq, N. Akhtar, H. Afzal, S. Khalid, M.R. Mufti, S. Hussain, A. Habib, G. Ahmad, A novel co-training-based approach for the classification of mental illnesses using social media posts, IEEE Access 7 (2019) 166165–166172, https://doi.org/10.1109/ACCESS.2019.2953087.
[86] G. Ranco, I. Bordino, G. Bormetti, G. Caldarelli, F. Lillo, M. Treccani, Coupling news sentiment with web browsing data improves prediction of intra-day price dynamics, PLoS ONE 11 (1) (2016) e0146576, https://doi.org/10.1371/journal.pone.0146576.
[87] J. Islam, Y. Zhang, Visual sentiment analysis for social images using transfer learning approach, in: 2016 IEEE International Conferences on Big Data and

Cloud Computing, Social Computing and Networking, Sustainable Computing and Communications, IEEE, 2016, pp. 124–130.
[88] N.H. Egebjerg, N. Hedegaard, G. Kuum, R.R. Mukkamala, R. Vatrapu, Big social data analytics in football: predicting spectators and TV ratings from Facebook data, in: 2017 IEEE International Congress on Big Data, BigData Congress, IEEE, 2017, pp. 81–88.
[89] W.T. Chu, Y.C. Chou, On broadcasted game video analysis: event detection, highlight detection, and highlight forecast, Multimed. Tools Appl. 76 (7) (2017) 9735–9758, https://doi.org/10.1007/s11042-016-3577-x.
[90] M.R. Bendre, R.C. Thool, V.R. Thool, Big data in precision agriculture: weather forecasting for future farming, in: 2015 1st International Conference on Next Generation Computing Technologies, IEEE, 2015, pp. 744–750.
[91] Y. Zhao, X. Xu, M. Wang, Predicting overall customer satisfaction: big data evidence from hotel online textual reviews, Int. J. Hosp. Manag. 76 (2019) 111–121, https://doi.org/10.1016/j.ijhm.2018.03.017.
[92] M.A. Kausar, V.S. Dhaka, S.K. Singh, Web crawler: a review, Int. J. Comput. Appl. 63 (2) (2013) 31–36, https://doi.org/10.5120/10440-5125.
[93] M.R. Murty, J.V.R. Murthy, P.R. PVGD, Text document classification based-on least square support vector machines with singular value decomposition, Int. J. Comput. Appl. 27 (7) (2011) 21–26, https://doi.org/10.5120/3312-4540.
[94] E. Boschee, P. Natarajan, R. Weischedel, Automatic extraction of events from open source text for predictive forecasting, in: Handbook of Computational Approaches to Counterterrorism, Springer, 2013, pp. 51–67.
[95] L. Getoor, A. Machanavajjhala, Entity resolution: theory, practice & open challenges, Proc. VLDB Endow. 5 (2012) 2018–2019, https://doi.org/10.14778/2367502.2367564.
[96] H. Köpcke, A. Thor, E. Rahm, Evaluation of entity resolution approaches on real-world match problems, Proc. VLDB Endow. 3 (2010) 484–493, https://doi.org/10.14778/1920841.1920904.
[97] R.K. Jena, Sentiment mining in a collaborative learning environment: capitalising on big data, Behav. Inf. Technol. 38 (9) (2019) 986–1001, https://doi.org/10.1080/0144929X.2019.1625440.
[98] S. Huston, J.S. Culpepper, W.B. Croft, Sketch-based indexing of n-words, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012, pp. 1864–1868.
[99] A.U. Rehman, A.K. Malik, B. Raza, W. Ali, A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis, Multimed. Tools Appl. 78 (18) (2019) 26597–26613, https://doi.org/10.1007/s11042-019-07788-7.
[100] L. Fan, Q. Wu, C. Ruan, Z. Zhuo, X. Wang, A feature extraction algorithm based on 2D complexity of Gabor wavelets transform for facial expression recognition, in: 2012 5th International Congress on Image and Signal Processing, IEEE, 2012, pp. 392–396.
[101] R. Muthukrishnan, M. Radha, Edge detection techniques for image segmentation, Int. J. Comput. Sci. Inf. Technol. 3 (6) (2012) 259.
[102] D. Jia, D. Wei, R. Socher, L.J. Li, L. Kai, F.F. Li, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[103] W. Xiang, C. Yang, L. Jiao, Q. Pei, Image content location privacy preserving in social network travel image sharing, in: International Conference on Artificial Intelligence and Security, Springer, 2020, pp. 617–628.
[104] D.W. Sun, C.J. Du, Segmentation of complex food images by stick growing and merging algorithm, J. Food Eng. 61 (1) (2004) 17–26, https://doi.org/10.1016/S0260-8774(03)00184-5.
[105] P.M. de Zeeuw, E. Ranguelova, E.J. Pauwels, Towards an online image-based tree taxonomy, in: 2007 7th Industrial Conference on Data Mining, Springer, 2007, pp. 296–306.
[106] W. Yun, D. Kim, C. Park, J. Kim, Hybrid facial representations for emotion recognition, ETRI J. 35 (6) (2013) 1021–1028, https://doi.org/10.4218/etrij.13.2013.0054.
[107] Y. Wang, H. Ren, Z. Qin, W. Zheng, L. Yu, Z. Geng, User context information prediction based on the mobile internet social pictures, in: 2016 2nd IEEE International Conference on Computer and Communications, IEEE, 2016, pp. 2397–2401.
[108] J. Cui, Z. Hao, H. Hu, S. Shan, X. Chen, Improving 2D face recognition via discriminative face depth estimation, in: 2018 International Conference on Biometrics, IEEE, 2018, pp. 140–147.
[109] R.L. Mandryk, M.V. Birk, The potential of game-based digital biomarkers for modeling mental health, JMIR Mental Heal. 6 (4) (2019) e13485, https://doi.org/10.2196/13485.
[110] M. Nilashi, A. Mardani, H. Liao, H. Ahmadi, A.A. Manaf, W. Almukadi, A hybrid method with TOPSIS and machine learning techniques for sustainable development of green hotels considering online reviews, Sustain. 11 (21) (2019) 6013, https://doi.org/10.3390/su11216013.
[111] R.Y.K. Lau, W. Zhang, W. Xu, Parallel aspect-oriented sentiment analysis for sales forecasting with big data, Prod. Oper. Manag. 27 (10) (2018) 1775–1794, https://doi.org/10.1111/poms.12737.
[112] S. Tanuwijaya, A. Alamsyah, M. Ariyanti, Mobile customer behaviour predictive analysis for targeting Netflix potential customer, in: 2021 9th International Conference on Information and Communication Technology, ICoICT, IEEE, 2021, pp. 348–352.
[113] T. Trzciński, P. Rokita, Predicting popularity of online videos using support vector regression, IEEE Trans. Multimed. 19 (2017) 2561–2570, https://doi.org/10.1109/TMM.2017.2695439.
[114] J. Chen, X. Song, L. Nie, X. Wang, H. Zhang, T.S. Chua, Micro tells macro: predicting the popularity of micro-videos via a transductive model, in: Proceedings of the 24th ACM International Conference on Multimedia, 2016, pp. 898–907.
[115] M.S. Rahim, A.Z.M.E. Chowdhury, M.A. Islam, M.R. Islam, Mining trailers data from youtube for predicting gross income of movies, in: 2017 IEEE Region 10 Humanitarian Technology Conference, R10-HTC, IEEE, 2017, pp. 551–554.
[116] S.K. Gaikwad, B.W. Gawali, P. Yannawar, A review on speech recognition technique, Int. J. Comput. Appl. 10 (2010) 16–24, https://doi.org/10.5120/1462-1976.
[117] N. Aggrawal, A. Arora, A. Anand, Y. Dwivedi, Early viewers or followers: a mathematical model for YouTube viewers' categorization, Kybernetes 50 (6) (2020) 1811–1836, https://doi.org/10.1108/K-03-2020-0128.
[118] V. da Silva, A.T. Winck, Video popularity prediction in data streams based on context-independent features, in: Proceedings of the Symposium on Applied Computing, 2017, pp. 95–100.
[119] W. Liu, Z. Duanmu, Z. Wang, End-to-end blind quality assessment of compressed videos using deep neural networks, in: ACM Multimedia, 2018, pp. 546–554.
[120] H. Dou, W.X. Zhao, Y. Zhao, D. Dong, J.-R. Wen, E.Y. Chang, Predicting the popularity of online content with knowledge-enhanced neural networks, in: ACM KDD, 2018.
[121] J.P. Verma, S. Agrawal, B. Patel, A. Patel, Big data analytics: challenges and applications for text, audio, video, and social media data, Int. J. Soft Comput. Artif. Intell. Appl. 5 (2016) 41–51, https://doi.org/10.5121/ijscai.2016.5105.
[122] A. Gandomi, M. Haider, Beyond the hype: Big data concepts, methods, and analytics, Int. J. Inf. Manag. 35 (2015) 137–144, https://doi.org/10.1016/j.ijinfomgt.2014.10.007.
[123] Y.B. Egorova, S.V. Skvortsova, L.V. Davydenko, Forecasting VT6 titanium alloy rolled bar mechanical properties, Metallurgist 64 (3) (2020) 242–252, https://doi.org/10.1007/s11015-020-00989-8.
[124] J. Wang, B. Zhang, Quality of environmental information disclosure and enterprise characteristics, Manag. Environ. Qual. An Int. J. 30 (2019) 963–979, https://doi.org/10.1108/MEQ-11-2018-0194.
[125] H. Wang, Y. Lu, C. Zhai, Latent aspect rating analysis on review text data: a rating regression approach, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2010, pp. 783–792.
[126] K. Song, F. Li, F. Long, J. Wang, Q. Ling, Discriminative deep feature learning for semantic-based image retrieval, IEEE Access 6 (2018) 44268–44280, https://doi.org/10.1109/ACCESS.2018.2862464.
[127] J. Li, L. Xu, L. Tang, S. Wang, L. Li, Big data in tourism research: a literature review, Tour. Manag. 68 (2018) 301–323, https://doi.org/10.1016/j.tourman.2018.03.009.
[128] I. Ahmed, M. Ahmad, G. Jeon, F. Piccialli, A framework for pandemic prediction using big data analytics, Big Data Res. 25 (2021) 100190, https://doi.org/10.1016/j.bdr.2021.100190.
[129] X. Hu, T. Oommen, Z. Lu, T. Wang, J.W. Kim, Consolidation settlement of Salt Lake County tailings impoundment revealed by time-series InSAR observations from multiple radar satellites, Remote Sens. Environ. 202 (2017) 199–209, https://doi.org/10.1016/j.rse.2017.05.023.
[130] Y.O. Sayad, H. Mousannif, H. Al Moatassime, Predictive modeling of wildfires: a new dataset and machine learning approach, Fire Saf. J. 104 (2019) 130–146, https://doi.org/10.1016/j.firesaf.2019.01.006.
[131] M.K. Saggi, S. Jain, Reference evapotranspiration estimation and modeling of the Punjab Northern India using deep learning, Comput. Electron. Agric. 156 (2019) 387–398, https://doi.org/10.1016/j.compag.2018.11.031.
[132] L. Wen, K. Zhou, S. Yang, X. Lu, Optimal load dispatch of community microgrid with deep learning based solar power and load forecasting, Energy 171 (15) (2019) 1053–1065, https://doi.org/10.1016/j.energy.2019.01.075.
[133] W. Jiang, H. Tang, L. Wu, H. Huang, H. Qi, Parallel processing of probabilistic models-based power supply unit mid-term load forecasting with apache spark, IEEE Access 7 (2019) 7588–7598, https://doi.org/10.1109/ACCESS.2018.2890339.
[134] J. Li, Q. He, H. Zhou, Y. Guan, W. Dai, Modeling driver behavior near intersections in hidden Markov model, Int. J. Environ. Res. Public Health 13 (12) (2016) 1265, https://doi.org/10.3390/ijerph13121265.
[135] K. Sultan, H. Ali, Z. Zhang, Call detail records driven anomaly detection and traffic prediction in mobile cellular networks, IEEE Access 6 (2018) 41728–41737, https://doi.org/10.1109/ACCESS.2018.2859756.
[136] Z. He, F. Yang, Z. Li, K. Liu, N. Xiong, Mining channel water depth information from IoT-based big automated identification system data for safe waterway navigation, IEEE Access 6 (2018) 75598–75608, https://doi.org/10.1109/ACCESS.2018.2883421.
[137] M.B. Arias, S. Bae, Electric vehicle charging demand forecasting model based on big data technologies, Appl. Energy 183 (2016) 327–339, https://doi.org/10.1016/j.apenergy.2016.08.080.

[138] X. Wu, Y. Lu, Y. Lin, Y. Yang, Measuring the destination accessibility of cycling transfer trips in metro station areas: a big data approach, Int. J. Environ. Res. Public Health 16 (15) (2019) 2641, https://doi.org/10.3390/ijerph16152641.

[139] R. Orus Perez, Using tensorflow-based neural network to estimate GNSS single frequency ionospheric delay (IONONet), Adv. Space Res. 63 (5) (2019) 1607–1618, https://doi.org/10.1016/j.asr.2018.11.011.

[140] L. Birek, A. Grzywaczewski, R. Iqbal, F. Doctor, V. Chang, A novel big data analytics and intelligent technique to predict driver's intent, Comput. Ind. 99 (2018) 226–240, https://doi.org/10.1016/j.compind.2018.03.025.

[141] Q. Lv, Y. Qiao, N. Ansari, J. Liu, J. Yang, Big data driven hidden Markov model based individual mobility prediction at points of interest, IEEE Trans. Veh. Technol. 66 (6) (2017) 5204–5216, https://doi.org/10.1109/TVT.2016.2611654.

[142] S. Jade, M.S.M. Vijayan, V.K. Gaur, T.P. Prabhu, S.C. Sahu, Estimates of precipitable water vapour from GPS data over the Indian subcontinent, J. Atmos. Sol.-Terr. Phys. 67 (6) (2005) 623–635, https://doi.org/10.1016/j.jastp.2004.12.010.

[143] K.G.S. Dharmawardana, J.N. Lokuge, P.S.B. Dassanayake, M.L. Sirisena, M.L. Fernando, A.S. Perera, S. Lokanathan, Predictive model for the dengue incidences in Sri Lanka using mobile network big data, in: 2017 IEEE International Conference on Industrial and Information Systems, IEEE, 2018, pp. 1–6.

[144] V. Cortés, J. Blasco, N. Aleixos, S. Cubero, P. Talens, Visible and near-infrared diffuse reflectance spectroscopy for fast qualitative and quantitative assessment of nectarine quality, Food Bioprocess Technol. 10 (10) (2017) 1755–1766, https://doi.org/10.1007/s11947-017-1943-y.

[145] H.M. Zawbaa, S. Schiano, L. Perez-Gandarillas, C. Grosan, A. Michrafy, C.Y. Wu, Computational intelligence modelling of pharmaceutical tabletting processes using bio-inspired optimization algorithms, Adv. Powder Technol. 29 (12) (2018) 2966–2977, https://doi.org/10.1016/j.apt.2018.11.008.

[146] S. Chatterjee, B. Datta, N. Dey, Hybrid neural network based rainfall prediction supported by flower pollination algorithm, Neural Netw. World 28 (6) (2018) 497–510, https://doi.org/10.14311/NNW.2018.28.027.

[147] F. Theuer, M.F. van Dooren, L. von Bremen, M. Kühn, Minute-scale power forecast of offshore wind turbines using long-range single-Doppler lidar measurements, Wind Energy Sci. 5 (4) (2020) 1449–1468, https://doi.org/10.5194/wes-5-1449-2020.

[148] L. Yang, X. Gao, J. Hua, P. Wu, Z. Li, D. Jia, Very short-term surface solar irradiance forecasting based on FengYun-4 geostationary satellite, Sensors 20 (9) (2020) 2606, https://doi.org/10.3390/s20092606.

[149] D.C. Yacchirema, D. Sarabia-Jacome, C.E. Palau, M. Esteve, A smart system for sleep monitoring by integrating IoT with big data analytics, IEEE Access 6 (2018) 35988–36001, https://doi.org/10.1109/ACCESS.2018.2849822.

[150] J. Xu, J. Meng, L.J. Quackenbush, Use of remote sensing to predict the optimal harvest date of corn, Field Crops Res. 236 (2019) 1–13, https://doi.org/10.1016/j.fcr.2019.03.003.

[151] S.M. Guzman, J.O. Paz, M.L.M. Tagert, The use of NARX neural networks to forecast daily groundwater levels, Water Resour. Manag. 31 (5) (2017) 1591–1603, https://doi.org/10.1007/s11269-017-1598-5.

[152] J.G. Hernandez-Travesio, C.M. Travesio, J.B. Alonso, M.K. Dutta, Applying data normalization for the solar radiation modelling, in: 2015 11th International Conference on Energy, Environment, Ecosystems and Sustainable Development, 2015, pp. 134–139.

[153] E. Habyarimana, I. Piccard, M. Catellani, P. De Franceschi, M. Dall'Agata, Towards predictive modeling of sorghum biomass yields using fraction of absorbed photosynthetically active radiation derived from sentinel-2 satellite imagery and supervised machine learning techniques, Agronomy 9 (4) (2019) 203, https://doi.org/10.3390/agronomy9040203.

[154] M. Sobhani, T. Hong, C. Martin, Temperature anomaly detection for electric load forecasting, Int. J. Forecast. 36 (2) (2020) 324–333, https://doi.org/10.1016/j.ijforecast.2019.04.022.

[155] A.H. Rabie, S.H. Ali, H.A. Ali, A.I. Saleh, A fog based load forecasting strategy for smart grids using big electrical data, Clust. Comput. 22 (1) (2019) 241–270, https://doi.org/10.1007/s10586-018-2848-x.

[156] S. Goudarzi, M.H. Anisi, N. Kama, F. Doctor, S.A. Soleymani, A.K. Sangaiah, Predictive modelling of building energy consumption based on a hybrid nature-inspired optimization algorithm, Energy Build. 196 (2019) 83–93, https://doi.org/10.1016/j.enbuild.2019.05.031.

[157] Y. Huang, L. Qian, A. Feng, N. Yu, Y. Wu, Short-term traffic prediction by two-level data driven model in 5G-enabled edge computing networks, IEEE Access 7 (2019) 123981–123991, https://doi.org/10.1109/ACCESS.2019.2938236.

[158] A. Wibisono, W. Jatmiko, H.A. Wisesa, B. Hardjono, P. Mursanto, Traffic big data prediction and visualization using fast incremental model trees-drift detection (FIMT-DD), Knowl.-Based Syst. 93 (2016) 33–46, https://doi.org/10.1016/j.knosys.2015.10.028.

[159] G. Thapa, K. Sharma, M.K. Ghose, Moving object detection and segmentation using frame differencing and summing technique, Int. J. Comput. Appl. 102 (7) (2014) 20–25, https://doi.org/10.5120/17828-8647.

[160] P. Khumprom, N. Yodo, A data-driven predictive prognostic model for lithium-ion batteries based on a deep learning algorithm, Energies 12 (4) (2019) 660, https://doi.org/10.3390/en12040660.

[161] A. Huang, Y. Huo, J. Yang, G. Li, Computational simulation and prediction on electrical conductivity of oxide-based melts by big data mining, Materials 12 (7) (2019) 1059, https://doi.org/10.3390/ma12071059.

[162] A. Ettehadtavakkol, A. Jamali, A data analytic workflow to forecast produced water from Marcellus shale, J. Nat. Gas Sci. Eng. 61 (2019) 293–302, https://doi.org/10.1016/j.jngse.2018.11.021.

[163] H. Asri, H. Mousannif, H. Al Moatassime, Real-time miscarriage prediction with SPARK, Proc. Comput. Sci. 113 (2017) 423–428, https://doi.org/10.1016/j.procs.2017.08.272.

[164] H. Wang, H. Su, STAR: a concise deep learning framework for citywide human mobility prediction, in: 2019 20th IEEE International Conference on Mobile Data Management, IEEE, 2019, pp. 304–309.

[165] S. Gan, S. Liang, K. Li, J. Deng, T. Cheng, Ship trajectory prediction for intelligent traffic management using clustering and ANN, in: 2016 UKACC 11th International Conference on Control, IEEE, 2016, pp. 1–6.

[166] T. Liu, G. Zhao, H. Wang, X. Hou, X. Qian, T. Hou, Finding optimal meteorological observation locations by multi-source urban big data analysis, in: 2016 7th International Conference on Cloud Computing and Big Data, IEEE, 2016, pp. 175–180.

[167] X. Wang, Z. Zhou, F. Xiao, K. Xing, Z. Yang, Y. Liu, C. Peng, Spatio-temporal analysis and prediction of cellular traffic in metropolis, IEEE Trans. Mob. Comput. 18 (9) (2018) 2190–2202, https://doi.org/10.1109/TMC.2018.2870135.

[168] Q. Wu, H. Lin, A novel optimal-hybrid model for daily air quality index prediction considering air pollutant factors, Sci. Total Environ. 683 (2019) 808–821, https://doi.org/10.1016/j.scitotenv.2019.05.288.

[169] S. Al-Janabi, M. Mohammad, A. Al-Sultan, A new method for prediction of air pollution based on intelligent computation, Soft Comput. 24 (1) (2020) 661–680, https://doi.org/10.1007/s00500-019-04495-1.

[170] C.J. Huang, P.-H. Kuo, A deep CNN-LSTM model for particulate matter ($PM_{2.5}$) forecasting in smart cities, Sensors 18 (7) (2018) 2220, https://doi.org/10.3390/s18072220.

[171] M. Gao, G. Shi, S. Li, Online prediction of ship behavior with automatic identification system sensor data using bidirectional long short-term memory recurrent neural network, Sensors 18 (12) (2018) 4211, https://doi.org/10.3390/s18124211.

[172] D.J. Gagne, A. McGovern, S.E. Haupt, J.K. Williams, Evaluation of statistical learning configurations for gridded solar irradiance forecasting, Sol. Energy 150 (2017) 383–393, https://doi.org/10.1016/j.solener.2017.04.031.

[173] A.Y. Saber, A.K.M.R. Alam, Short term load forecasting using multiple linear regression for big data, in: 2017 IEEE Symposium Series on Computational Intelligence, IEEE, 2017, pp. 1–6.

[174] S. Haben, G. Giasemidis, F. Ziel, S. Arora, Short term load forecasting and the effect of temperature at the low voltage level, Int. J. Forecast. 35 (4) (2019) 1469–1484, https://doi.org/10.1016/j.ijforecast.2018.10.007.

[175] S. Haben, G. Giasemidis, A hybrid model of kernel density estimation and quantile regression for GEFCom2014 probabilistic load forecasting, Int. J. Forecast. 32 (3) (2016) 1017–1022, https://doi.org/10.1016/j.ijforecast.2015.11.004.

[176] F. Orsini, M. Gastaldi, L. Mantecchini, R. Rossi, Neural networks trained with WiFi traces to predict airport passenger behavior, in: 2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems, IEEE, 2019, pp. 1–7.

[177] A. Emami, M. Sarvi, S.A. Bagloee, Short-term traffic flow prediction based on faded memory Kalman Filter fusing data from connected vehicles and Bluetooth sensors, Simul. Model. Pract. Theory 102 (2020) 102025, https://doi.org/10.1016/j.simpat.2019.102025.

[178] L. Wen, K. Zhou, S. Yang, L. Li, Compression of smart meter big data: a survey, Renew. Sustain. Energy Rev. 91 (2018) 59–69, https://doi.org/10.1016/j.rser.2018.03.088.

[179] X. Guo, Application of meteorological big data, in: 2016 16th International Symposium on Communications and Information Technologies, IEEE, 2016, pp. 273–279.

[180] N. Sun, J. Zhou, L. Chen, B. Jia, M. Tayyab, T. Peng, An adaptive dynamic short-term wind speed forecasting model using secondary decomposition and an improved regularized extreme learning machine, Energy 165 (2018) 939–957, https://doi.org/10.1016/j.energy.2018.09.180.

[181] H. Naganathan, W.K. Chong, Z. Huang, Y. Cheng, A non-stationary analysis using ensemble empirical mode decomposition to detect anomalies in building energy consumption, Proc. Eng. 145 (2016) 1059–1065, https://doi.org/10.1016/j.proeng.2016.04.137.

[182] S.J. Al'Aref, G. Singh, A.R. van Rosendael, K.K. Kolli, X. Ma, G. Maliakal, M. Pandey, B.C. Lee, J. Wang, Z. Xu, Y. Zhang, J.K. Min, S.C. Wong, R.M. Minutello, Determinants of in-hospital mortality after percutaneous coronary intervention: a machine learning approach, J. Am. Heart Assoc. 8 (5) (2019) e011160, https://doi.org/10.1161/JAHA.118.011160.

[183] P. Smith, Google's Midas touch: predicting UK unemployment with internet search data, J. Forecast. 35 (2016) 263–284, https://doi.org/10.1002/for.2391.

[184] S. Kim, D.H. Shin, Forecasting short-term air passenger demand using big data from search engine queries, Autom. Constr. 70 (2016) 98–108, https://doi.org/10.1016/j.autcon.2016.06.009.

[185] M. Wen, P. Li, L. Zhang, Y. Chen, Stock market trend prediction using high-order information of time series, IEEE Access 7 (2019) 28299–28308, https://doi.org/10.1109/ACCESS.2019.2901842.
[186] M. Jena, R.K. Behera, S.K. Rath, Machine learning models for stock prediction using real-time streaming data, in: International Conference on Biologically Inspired Techniques in Many-Criteria Decision Making, Springer, 2019, pp. 101–108.
[187] J. Wang, R. Hou, C. Wang, L. Shen, Improved v-support vector regression model based on variable selection and brain storm optimization for stock price forecasting, Appl. Soft Comput. J. 49 (2016) 164–178, https://doi.org/10.1016/j.asoc.2016.07.024.
[188] M. Inthachot, V. Boonjing, S. Intakosum, Artificial neural network and genetic algorithm hybrid intelligence for predicting Thai stock price index trend, Comput. Intell. Neurosci. 15 (2016) 1–8, https://doi.org/10.1155/2016/3045254.
[189] X. Liu, J. Li, Using support vector machine for online purchase predication, in: 2016 International Conference on Logistics, Informatics and Service Sciences, IEEE, 2016, pp. 1–6.
[190] J. Liang, J. Yang, Y. Wu, C. Li, L. Zheng, Big data application in education: dropout prediction in edx MOOCs, in: 2016 IEEE Second International Conference on Multimedia Big Data, IEEE, 2016, pp. 440–443.
[191] C. Urkup, B. Bozkaya, F. Sibel Salman, Customer mobility signatures and financial indicators as predictors in product recommendation, PLoS ONE 13 (7) (2018) e0201197, https://doi.org/10.1371/journal.pone.0201197.
[192] S.K.S. Fan, C.J. Su, H.T. Nien, P.F. Tsai, C.Y. Cheng, Using machine learning and big data approaches to predict travel time based on historical and real-time data from Taiwan electronic toll collection, Soft Comput. 22 (17) (2018) 5707–5718, https://doi.org/10.1007/s00500-017-2610-y.
[193] Y. Algan, F. Murtin, E. Beasley, K. Higa, C. Senik, Well-being through the lens of the Internet, PLoS ONE 14 (1) (2019) e0209562, https://doi.org/10.1371/journal.pone.0209562.
[194] C. Anagnostopoulos, F. Savva, P. Triantafillou, Scalable aggregation predictive analytics, Appl. Intell. 48 (9) (2018) 2546–2567, https://doi.org/10.1007/s10489-017-1093-y.
[195] J.C. Eichstaedt, R.J. Smith, R.M. Merchant, L.H. Ungar, P. Crutchley, D. Preoțiuc-Pietro, D.A. Asch, H.A. Schwartz, Facebook language predicts depression in medical records, Proc. Natl. Acad. Sci. USA 115 (44) (2018) 11203–11208, https://doi.org/10.1073/pnas.1802331115.
[196] Y. Zhang, E. Yao, J. Zhang, K. Zheng, Estimating metro passengers' path choices by combining self-reported revealed preference and smart card data, Transp. Res., Part C, Emerg. Technol. 92 (2018) 76–89, https://doi.org/10.1016/j.trc.2018.04.019.
[197] J. Henriques, L. Bernardo, R. Oliveira, P. Amaral, F. Ganhao, P. Pinto, R. Dinis, Outliers detection in network services with self-learned profiles, in: 2017 9th International Congress on Ultra-Modern Telecommunications and Control Systems and Workshops, IEEE, 2017, pp. 238–243.
[198] T.M. Deist, A. Jochems, J. van Soest, G. Nalbantov, C. Oberije, S. Walsh, M. Eble, P. Bulens, P. Coucke, W. Dries, A. Dekker, P. Lambin, Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT, Clin. Transl. Radiat. Oncol. 4 (2017) 24–31, https://doi.org/10.1016/j.ctro.2016.12.004.
[199] K. Jakhar, R. Bajaj, R. Gupta, Pneumothorax segmentation: deep learning image segmentation to predict pneumothorax, arXiv:1912.07329, 2019.
[200] Q. Liang, Y. Nan, G. Coppola, K. Zou, W. Sun, D. Zhang, Y. Wang, G. Yu, Weakly supervised biomedical image segmentation by reiterative learning, IEEE J. Biomed. Health Inform. 23 (3) (2019) 1205–1214, https://doi.org/10.1109/JBHI.2018.2850040.
[201] J. Wang, C. Li, Y. Chen, X. Ji, Y. Liu, H. Zhang, P. Shi, S. Zhang, Automatic filter of normal papanicolaou smear using multi-instance learning algorithms, in: 2015 10th International Conference on Intelligent Systems and Knowledge Engineering, IEEE, 2015, pp. 420–423.
[202] E. Pashaei, N. Aydin, Markovian encoding models in human splice site recognition using SVM, Comput. Biol. Chem. 73 (2018) 159–170, https://doi.org/10.1016/j.compbiolchem.2018.02.005.
[203] R.A. Taylor, J.R. Pare, A.K. Venkatesh, H. Mowafi, E.R. Melnick, W. Fleischman, M.K. Hall, Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach, Acad. Emerg. Med. 23 (3) (2016) 269–278, https://doi.org/10.1111/acem.12876.
[204] S.T. Prasad, S. Sangavi, A. Deepa, F. Sairabanu, R. Ragasudha, Diabetic data analysis in big data with predictive method, in: 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies, IEEE, 2017, pp. 1–4.
[205] H. Farhidzadeh, D.B. Goldgof, L.O. Hall, R.A. Gatenby, R.J. Gillies, M. Raghavan, Texture feature analysis to predict metastatic and necrotic soft tissue sarcomas, in: 2015 IEEE International Conference on Systems, Man, and Cybernetics, IEEE, 2015, pp. 2798–2802.
[206] H.R. Hassanzadeh, J.H. Phan, M.D. Wang, A multi-modal graph-based semi-supervised pipeline for predicting cancer survival, in: 2016 IEEE International Conference on Bioinformatics and Biomedicine, IEEE, 2016, pp. 184–189.
[207] F.A. Batarseh, E.A. Latif, Assessing the quality of service using big data analytics: with application to healthcare, Big Data Res. 4 (2016) 13–24, https://doi.org/10.1016/j.bdr.2015.10.001.
[208] X. Xu, Y. Cui, Forecasting on equipment manufacturing industry development in view of big data, in: 2017 International Conference on Applied Mathematics, Modelling and Statistics Application, AMMSA 2017, Atlantis Press, 2017, pp. 266–269.
[209] L. Xiaoxuan, W. Qi, P. Geng, L. Benfu, Tourism forecasting by search engine data with noise-processing, Afr. J. Bus. Manag. 10 (6) (2016) 114–130, https://doi.org/10.5897/AJBM2015.7945.
[210] J. Tang, Evaluation of the forecast models of Chinese tourists to Thailand based on search engine attention: a case study of Baidu, Wirel. Pers. Commun. 102 (4) (2018) 3825–3833, https://doi.org/10.1007/s11277-018-5413-2.
[211] W. Lu, H. Rui, C. Liang, L. Jiang, S. Zhao, K. Li, A method based on GA-CNN-LSTM for daily tourist flow prediction at scenic spots, Entropy 22 (3) (2020) 261, https://doi.org/10.3390/e22030261.
[212] Z. Wang, Y. Huang, B. He, T. Luo, Y. Wang, Y. Fu, Short-term infectious diarrhea prediction using weather and search data in Xiamen, China, Sci. Program. 2020 (2020) 1–12, https://doi.org/10.1155/2020/8814222.
[213] M. Verma, K. Kishore, M. Kumar, A.R. Sondh, G. Aggarwal, S. Kathirvel, Google search trends predicting disease outbreaks: an analysis from India, Healthc. Inform. Res. 24 (4) (2018) 300–308, https://doi.org/10.4258/hir.2018.24.4.300.
[214] L. Tang, C. Zhang, L. Li, S. Wang, A multi-scale method for forecasting oil price with multi-factor search engine data, Appl. Energy 257 (2020) 114033, https://doi.org/10.1016/j.apenergy.2019.114033.
[215] V.P. Ramesh, P. Baskaran, A. Krishnamoorthy, D. Damodaran, P. Sadasivam, Back propagation neural network based big data analytics for a stock market challenge, Commun. Stat., Theory Methods 48 (14) (2019) 3622–3642, https://doi.org/10.1080/03610926.2018.1478103.
[216] A. Nahil, A. Lyhyaoui, Short-term stock price forecasting using kernel principal component analysis and support vector machines: the case of Casablanca stock exchange, Proc. Comput. Sci. 127 (2018) 161–169, https://doi.org/10.1016/j.procs.2018.01.111.
[217] S. Jeon, B. Hong, H. Lee, J. Kim, Stock price prediction based on stock big data and pattern graph analysis, in: International Conference on Internet of Things and Big Data, SCITEPRESS, 2016, pp. 223–231.
[218] S. Carta, A. Medda, A. Pili, D.R. Recupero, R. Saia, Forecasting e-commerce products prices by combining an autoregressive integrated moving average (ARIMA) model and Google trends data, Future Internet 11 (1) (2019) 5, https://doi.org/10.3390/fi11010005.
[219] K. Fang, Y. Jiang, M. Song, Customer profitability forecasting using big data analytics: a case study of the insurance industry, Comput. Ind. Eng. 101 (2016) 554–564, https://doi.org/10.1016/j.cie.2016.09.011.
[220] N.S. Arunraj, D. Ahrens, M. Fernandes, Application of SARIMAX model to forecast daily sales in food retail industry, Int. J. Oper. Res. Inf. Syst. 7 (2) (2016) 1–21, https://doi.org/10.4018/IJORIS.2016040101.
[221] F. Yoseph, M. Heikkilä, D. Howard, Outliers identification model in point-of-sales data using enhanced normal distribution method, in: 2019 International Conference on Machine Learning and Data Engineering, IEEE, 2019, pp. 72–78.
[222] A. Goyal, S. Krishnamurthy, S. Kulkarni, R. Kumar, M. Vartak, M.A. Lanham, A solution to forecast demand using long short-term memory recurrent neural networks for time series forecasting, in: Midwest Decision Sciences Institute Conference, IEEE, 2018, pp. 1–18.
[223] L. Cen, D. Ruta, A map-based gender prediction model for big e-commerce data, in: 2017 IEEE International Conference on Internet of Things (IThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), IEEE, 2017, pp. 1025–1029.
[224] B. Zheng, B. Liu, A scalable purchase intention prediction system using extreme gradient boosting machines with browsing content entropy, in: 2018 IEEE International Conference on Consumer Electronics, IEEE, 2018, pp. 1–4.
[225] Q. Li, S. Meng, S. Zhang, M. Wu, J. Zhang, M. Taleby Ahvanooey, M.S. Aslam, Safety risk monitoring of cyber-physical power systems based on ensemble learning algorithm, IEEE Access 7 (2019) 24788–24805, https://doi.org/10.1109/ACCESS.2019.2896129.
[226] S. Benabderrahmane, N. Mellouli, M. Lamolle, Predicting the users' click-streams using time series representation, symbolic sequences, and deep learning: application on job offers recommendation tasks, in: 2017 IEEE International Conference on Information Reuse and Integration, IEEE, 2017, pp. 436–443.
[227] P. Liu, Y. Zhang, D. Kong, B. Yin, Improved spatio-temporal residual networks for bus traffic flow prediction, Appl. Sci. 9 (4) (2019) 615, https://doi.org/10.3390/app9040615.
[228] Y. Zhang, L. Yakob, M.B. Bonsall, W. Hu, Predicting seasonal influenza epidemics using cross-hemisphere influenza surveillance data and local internet query data, Sci. Rep. 9 (1) (2019) 1–7, https://doi.org/10.1038/s41598-019-39871-2.
[229] W.B. Nicholson, D.S. Matteson, J. Bien, VARX-L: structured regularization for large vector autoregressions with exogenous variables, Int. J. Forecast. 33 (3) (2017) 627–651, https://doi.org/10.1016/j.ijforecast.2017.01.003.
[230] X. Yang, Y. Tong, X. Meng, S. Zhao, Z. Xu, Y. Li, X. Jia, S. Tan, Adaptive logistic group Lasso method for predicting the no-reflow among the multiple types of high-dimensional variables with missing data, in: 2016 7th IEEE Interna-

tional Conference on Software Engineering and Service Science, IEEE, 2016, pp. 1085–1089.
[231] C. Poirier, A. Lavenu, V. Bertaud, B. Campillo-Gimenez, E. Chazard, M. Cuggia, G. Bouzillé, Real time influenza monitoring using hospital big data in combination with machine learning methods: comparison study, JMIR Public Heal. Surveill. 4 (4) (2018) e11361, https://doi.org/10.2196/11361.
[232] L. Chen, Z. Qiao, M. Wang, C. Wang, R. Du, H.E. Stanley, Which artificial intelligence algorithm better predicts the Chinese stock market?, IEEE Access 6 (2018) 48625–48633, https://doi.org/10.1109/ACCESS.2018.2859809.
[233] T. Yang, K. Qian, D.C.T. Lo, Y. Xie, Y. Shi, L. Tao, Improve the prediction accuracy of naïve Bayes classifier with association rule mining, in: 2016 IEEE 2nd International Conference on Big Data Security on Cloud, IEEE, 2016, pp. 129–133.
[234] B. Wang, P. Liu, Z. Chao, W. Junmei, W. Chen, N. Cao, G.M.P. O'Hare, F. Wen, Research on hybrid model of garlic short-term price forecasting based on big data, Comput. Mater. Contin. 57 (2) (2018) 283–296, https://doi.org/10.32604/cmc.2018.03791.
[235] L. Ismail, H. Materwala, A.P. Karduck, A. Adem, Requirements of health data management systems for biomedical care and research: scoping review, J. Med. Internet Res. 22 (7) (2020) e17508, https://doi.org/10.2196/17508.
[236] F.R. Lucini, F.S. Fogliatto, G.J.C. da Silveira, J.L. Neyeloff, M.J. Anzanello, R.S. Kuchenbecker, B.D. Schaan, Text mining approach to predict hospital admissions using early medical records from the emergency department, Int. J. Med. Inform. 100 (2017) 1–8, https://doi.org/10.1016/j.ijmedinf.2017.01.001.
[237] C. Meng, L. Zhou, Big data encryption technology based on ASCII and application on credit supervision, in: 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering, IEEE, 2020, pp. 79–82.
[238] L. Xu, G. Shao, Y. Cao, H. Yang, C. Sun, T. Zhang, B. Wen, X. Cheng, C. Song, X. He, Research on telecom big data platform of LTE/5G mobile networks, in: 2019 IEEE International Conferences on Ubiquitous Computing & Communications, IEEE, 2019, pp. 756–761.
[239] Y. Li, N. Li, Y. Xia, Research on a data desensitization algorithm of blockchain distributed energy transaction based on differential privacy, in: 2019 IEEE 8th International Conference on Advanced Power System Automation and Protection, IEEE, 2019, pp. 980–985.
[240] M. Castellanos, B. Zhang, I. Jimenez, P. Ruiz, M. Durazo, U. Dayal, L. Jow, Data desensitization of customer data for use in optimizer performance experiments, in: 2010 IEEE 26th International Conference on Data Engineering, ICDE 2010, IEEE, 2010, pp. 1081–1092.