Documenting and predicting topic changes in Computers in Biology and Medicine: A bibliometric keyword analysis from 1990 to 2017

FAUST, Oliver <http://orcid.org/0000-0002-0352-6716>

Available from Sheffield Hallam University Research Archive (SHURA) at:
http://shura.shu.ac.uk/18989/

This document is the author deposited version. You are advised to consult the publisher's version if you wish to cite from it.

Published version


Copyright and re-use policy

See http://shura.shu.ac.uk/information.html
Documenting and predicting topic changes in Computers in Biology and Medicine: A bibliometric keyword analysis from 1990 to 2017

Oliver Faust

Department of Engineering and Mathematics, Sheffield Hallam University, UK

ARTICLE INFO

Keywords:
Research trends
Topic analysis
Topic detection and tracking
Text mining
Computers in biology and medicine

ABSTRACT

The Computers in Biology and Medicine (CBM) journal promotes the use of computing machinery in the fields of bioscience and medicine. Since the first volume in 1970, the importance of computers in these fields has grown dramatically, this is evident in the diversification of topics and an increase in the publication rate. In this study, we quantify both change and diversification of topics covered in. This is done by analysing the author supplied keywords, since they were electronically captured in 1990. The analysis starts by selecting 40 keywords, related to Medical (M) (7), Data (D) (10), Feature (F) (17) and (AI) (6) methods. Automated keyword clustering shows the statistical connection between the selected keywords. We found that the three most popular topics in CBM are: Support Vector Machine (SVM), Electroencephalography (EEG) and IMAGE PROCESSING. In a separate analysis step, we bagged the selected keywords into sequential one year time slices and calculated the normalized appearance. The results were visualised with graphs that indicate the CBM topic changes. These graphs show that there was a transition from Artificial Neural Network (ANN) to SVM. In 2006 SVM replaced ANN as the most important AI algorithm. Our investigation helps the editorial board to manage and embrace topic change. Furthermore, our analysis is interesting for the general reader, as the results can help them to adjust their research directions.

1. Introduction

Documenting the use of computers in bioscience and medicine is a very dynamic endeavour. Therefore, Computers in Biology and Medicine (CBM) is a journal which was set-up as a forum to publish scientific articles and reviews. The content areas include medical disease diagnosis [1–4], medical data [5], information processing [6–8] and dissemination [9]. Medical disease creates the need to build physical problem solutions and computer methods realize the required functionality [10,11]. The problem solutions can take the form of biochemical [12], biocontrol [13], neural simulation [14] and automatic computer analysis systems [15–17]. Keeping track of topic changes in that scientific area is important for steering the use of computing machinery in medicine and biology towards novel and forward thinking applications. However, the diverse and dynamic nature of the forum makes it difficult to track and analyse topics over time.

Bibliographic research aims to provide an overview of trends and issues encountered in dynamic literature [18–20]. As such, it is a meta-analysis method which is applied to a substantial body of research literature [21,22]. Not only the topics and the writing style evolve, also the bibliographic features of the documents change over time. For example, the author supplied keywords in CBM are only captured since 1990. Before that, Topic Detection and Tracking (TDT) [23] was difficult and error prone, because a third party had to extract the topic from the paper text. The ability to conduct TDT studies is significant for meta research on science, technology and policy [24,25]. TDT tools can be used to profile research [26,27], document trends and topics [28–30] as well as analyse research impact [31,32]. Linear models can be used to predict incremental change, but they underperform when it comes to predicting disruptive and revolutionary events [33]. Unfortunately, traditional methods use linear models on static data [34]. For example, keyword cluster analysis is static, because the clusters do not reflect change over time. Therefore, these methods cannot be used to document and predict topic changes [33].

To address the issues raised above, we have analysed the topics covered in CBM with static and dynamic keyword analysis methods. We have applied static frequency and cluster analysis to author supplied keywords from all papers published in CBM since 1990. The frequency analysis shows that Support Vector Machine (SVM), Electroencephalography (EEG), and IMAGE PROCESSING are the most widely used keywords in CBM publications. The cluster analysis reveals the structure within the keyword co-occurrence matrix. With a second analysis step,
we found the normalized appearance of the most important keywords in yearly intervals from 1990 to 2017. Analysing the normalized keyword appearance reveals the topic change dynamics. In order to interpret these dynamics, we put forward that future topics in CBM will be influenced by Medical (M) needs and advances in Artificial Intelligence (AI), Feature (F) extraction and medical Data (D) acquisition. Based on these four categories, we found that the most striking topic change happened in AI, namely the transition from Artificial Neural Network (ANN) to SVM. The static and dynamic bibliographic analysis results can serve as a basis for the editorial board to keep CBM relevant for the advancement of science and technology. The bibliometric research results are interesting for the general reader as well, because they reveal both statistical connections between keywords and trending topics. That information is useful when it comes to deciding on what technology to focus.

**Table 1**

<table>
<thead>
<tr>
<th>Keyword Correctiondone to the original author keywords.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Keyword</strong></td>
</tr>
<tr>
<td>ALGORITHMS</td>
</tr>
<tr>
<td>SUPPORT VECTOR MACHINE</td>
</tr>
<tr>
<td>SUPPORT VECTOR MACHINE (SVM)</td>
</tr>
<tr>
<td>SVM</td>
</tr>
<tr>
<td>UERAL NETWORKS</td>
</tr>
<tr>
<td>UERAL NETWORK</td>
</tr>
<tr>
<td>INITE ELEMENT METHOD</td>
</tr>
<tr>
<td>INITE ELEMENT ANALYSIS</td>
</tr>
<tr>
<td>ELECTROCARDIOGRAM</td>
</tr>
<tr>
<td>ELECTROCARDIOGRAM (ECG)</td>
</tr>
<tr>
<td>ELECTROENCEPHALOGRAM</td>
</tr>
<tr>
<td>ELECTROENCEPHALOGRAM (EEG)</td>
</tr>
<tr>
<td>MAGNETIC RESONANCE IMAGING</td>
</tr>
<tr>
<td>COMPUTED TOMOGRAPHY</td>
</tr>
<tr>
<td>ARTIFICIAL NEURAL NETWORK</td>
</tr>
<tr>
<td>ARTIFICIAL INTELLIGENCE</td>
</tr>
<tr>
<td>VIRTUAL REALITY</td>
</tr>
<tr>
<td>MODELING</td>
</tr>
<tr>
<td>HEART RATE VARIABILITY</td>
</tr>
<tr>
<td>AUTONOMOUS NERVOUS SYSTEM</td>
</tr>
<tr>
<td>GENETIC ALGORITHM</td>
</tr>
</tbody>
</table>

**Fig. 2.** Number of articles published in from 1990 (Volume 20) to 2017 (Volume 87).
The paper text refers to the bold part in Table 2: “The time slice 2015 (highlighted) links the example, shown in Equation (2), with the content of the SVM.csv file, shown in Table 2.

To substantiate our claim of supporting both editorial management and readers of CBM, we have organized the rest of this paper as follows. The next section introduces the materials used for the bibliographical research. Section 3 presents the research results. These results are discussed in Section 4. The discussion centres on interpreting and relating our findings to the wider research community. The study concludes with Section 5.

2. Materials

We had two distinct ideas for analysing the author supplied keywords in CBM papers. The first idea was keyword cluster analysis. There is good tool support for that method and the algorithms are well established. Cluster analysis reveals the statistical connections between individual topics [35]. However, cluster analysis does not document topic changes. With the second idea we address that shortcoming by incorporating the discussion process. Table 1 documents the information loss, with respect to our meta-analysis, for doing this first entry in Table 1 documents a change made during the refinement process. The information loss, with respect to our meta-analysis, for doing this first entry in Table 1 documents a change made during the refinement process. The information loss, with respect to our meta-analysis, for doing this first entry in Table 1 documents a change made during the refinement process. The information loss, with respect to our meta-analysis, for doing this first entry in Table 1 documents a change made during the refinement process. The information loss, with respect to our meta-analysis, for doing this first entry in Table 1 documents a change made during the refinement process. The information loss, with respect to our meta-analysis, for doing this first entry in Table 1 documents a change made during the refinement process. The information loss, with respect to our meta-analysis, for doing this first entry in Table 1 documents a change made during the refinement process.

The next section introduces the materials used for the bibliographical research. Section 3 presents the research results. These results are discussed in Section 4. The discussion centres on interpreting and relating our findings to the wider research community. The study concludes with Section 5.

2.1. Data acquisition

Data acquisition started with a straightforward publication search in the web of science [38]. The search term ‘Computers in Biology and Medicine’ resulted in a data set with 3406 entries. Each entry represents a paper that was published in the time period from 1970 to 2017. However, careful analysis reveals that the web of science captured author supplied keywords only from Volume 20 Issue 1 in 1990 onward. Restricting the publication years to the timespan from 1990 (Volume 20 Issue 1) to 2017 (Volume 91,2 published 1 December 2017) yielded 2946 entries. Fig. 2 shows the distribution of these entries. For example, the graph in Fig. 2 indicates that there were 311 articles published in 2015. The keywords of these papers became the raw data for the refinement process which is discussed in the next section.

2.2. Keyword refinement

The refinement process aims to extract the most relevant keywords. The process follows the flowchart shown in Fig. 1. It starts with an initial analysis, which revealed that the documents contained 15707 keywords and 10016 of these keywords were unique. Furthermore, the analysis indicated that the unique keywords were inconsistent. These inconsistencies arise from singular and plural forms of the same word as well as spelling differences. In order to correct the inconsistencies, we entered into an iterative refinement process. Table 1 documents the changes made during the refinement process. Most of these changes are self-evident. For example, the first entry in Table 1 documents a change from the plural form of the keyword ALGORITHM(S) to its singular form. The information loss, with respect to our meta-analysis, for doing this change is minimal. The acronyms were introduced to make the result more readable.

Once the keywords were consistent, we progressed to a second refinement step. This step removed irrelevant keywords. To facilitate this irrelevance reduction, we specified that each keyword must appear 15 or more times in the 2946 papers. 57 keywords met this criteria. However,

<table>
<thead>
<tr>
<th>Table 2</th>
<th>SVM.csv file content, where ‘na’ stands for ‘Normalized appearance in %’.</th>
</tr>
</thead>
<tbody>
<tr>
<td>year</td>
<td>na</td>
</tr>
<tr>
<td>1990</td>
<td>0</td>
</tr>
<tr>
<td>1991</td>
<td>0</td>
</tr>
<tr>
<td>1992</td>
<td>0</td>
</tr>
<tr>
<td>1993</td>
<td>0</td>
</tr>
<tr>
<td>1994</td>
<td>0</td>
</tr>
<tr>
<td>1995</td>
<td>0</td>
</tr>
<tr>
<td>1996</td>
<td>0</td>
</tr>
<tr>
<td>1997</td>
<td>0</td>
</tr>
<tr>
<td>1998</td>
<td>0</td>
</tr>
<tr>
<td>1999</td>
<td>0</td>
</tr>
<tr>
<td>2000</td>
<td>0</td>
</tr>
<tr>
<td>2001</td>
<td>2.7027</td>
</tr>
<tr>
<td>2002</td>
<td>0</td>
</tr>
<tr>
<td>2003</td>
<td>0</td>
</tr>
</tbody>
</table>

2 Since Volume 44, discontinued issues – only volumes are used.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Keywords mapped to the analysis results. C indicates the cluster number and m is the gradient value. The list is ordered in accordance with the cluster number. The individual cluster entries are ordered by the keyword occurrence.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyword</td>
<td>Area</td>
</tr>
<tr>
<td>IMAGE PROCESSING</td>
<td>F</td>
</tr>
<tr>
<td>SEGMENTATION</td>
<td>F</td>
</tr>
<tr>
<td>Magnetic Resonance Imaging (MRI)</td>
<td>D</td>
</tr>
<tr>
<td>Computed Tomography (CT)</td>
<td>D</td>
</tr>
<tr>
<td>IMAGE SEGMENTATION</td>
<td>F</td>
</tr>
<tr>
<td>IMAGE ANALYSIS</td>
<td>F</td>
</tr>
<tr>
<td>ULTRASOUND</td>
<td>D</td>
</tr>
<tr>
<td>MEDICAL IMAGING</td>
<td>D</td>
</tr>
<tr>
<td>IMAGE REGISTRATION</td>
<td>F</td>
</tr>
<tr>
<td>MYOCARDIAL INFARCTION</td>
<td>M</td>
</tr>
<tr>
<td>Electrocardiogram (ECG)</td>
<td>D</td>
</tr>
<tr>
<td>Heart Rate Variability (HRV)</td>
<td>D</td>
</tr>
<tr>
<td>ATRIAL FIBRILLATION</td>
<td>M</td>
</tr>
<tr>
<td>WAVELET TRANSFORM</td>
<td>F</td>
</tr>
<tr>
<td>SPECTRAL ANALYSIS</td>
<td>F</td>
</tr>
<tr>
<td>Principal Component Analysis (PCA)</td>
<td>F</td>
</tr>
<tr>
<td>ENTROPY</td>
<td>F</td>
</tr>
<tr>
<td>FRACTAL DIMENSION</td>
<td>F</td>
</tr>
<tr>
<td>Autonomic Nervous System (ANS)</td>
<td>M</td>
</tr>
<tr>
<td>SVM</td>
<td>AI</td>
</tr>
<tr>
<td>GENETIC ALGORITHM</td>
<td>AI</td>
</tr>
<tr>
<td>ANN</td>
<td>AI</td>
</tr>
<tr>
<td>Discrete Wavelet Transform (DWT)</td>
<td>F</td>
</tr>
<tr>
<td>CANCER</td>
<td>M</td>
</tr>
<tr>
<td>PHARMACOKINETICS</td>
<td>F</td>
</tr>
<tr>
<td>GENE EXPRESSION DATA</td>
<td>D</td>
</tr>
<tr>
<td>FUZZY LOGIC</td>
<td>AI</td>
</tr>
<tr>
<td>Neuroal Network (NN)</td>
<td>AI</td>
</tr>
<tr>
<td>BREAST CANCER</td>
<td>M</td>
</tr>
<tr>
<td>CLUSTERING</td>
<td>AI</td>
</tr>
<tr>
<td>GENE EXPRESSION</td>
<td>F</td>
</tr>
<tr>
<td>MAMMOGRAPHY</td>
<td>D</td>
</tr>
<tr>
<td>MICROARRAY DATA</td>
<td>D</td>
</tr>
<tr>
<td>Finite Element Method (FEM)</td>
<td>F</td>
</tr>
<tr>
<td>FLUID DYNAMICS</td>
<td>F</td>
</tr>
<tr>
<td>ATHEROSCLEROSIS</td>
<td>M</td>
</tr>
<tr>
<td>WALL SHEAR STRESS</td>
<td>F</td>
</tr>
<tr>
<td>BLOOD FLOW</td>
<td>F</td>
</tr>
<tr>
<td>ECG</td>
<td>D</td>
</tr>
<tr>
<td>EPILEPSY</td>
<td>M</td>
</tr>
</tbody>
</table>

| Total | 1117 | 78.93 ⋅ 10⁻² |
17 of these keywords were not relevant for documenting and predicting topic changes. The following list details the removed keywords and provides the rational as to why they were deemed to be irrelevant.

- CLASSIFICATION, COMPUTER SIMULATION, BIOINFORMATICS, MODELING, MODEL, ALGORITHM and SOFTWARE – These keywords are not focused enough to shed light on potential future trends in CBM.
- EXPERT SYSTEM, COMPUTER-AIDED DIAGNOSIS, DECISION SUPPORT, MACHINE LEARNING and DIAGNOSIS – These are important keyword categories. However, the importance of these categories for topic changes comes from the underlying methods.
- DATA MINING, SIGNAL PROCESSING and DATA ANALYSIS – The keywords are too broad. Most studies in CBM involve some sort of data analysis and signal processing. Data mining is a more recent term, however it is the underlying method which makes the topic important.
- PREDICTION – The term is covered by the AI methods.
- VISUALIZATION – Most of the CBM studies involve some sort of visualization. Therefore, this term is not a good indicator of future trends.
- OPTIMIZATION – As such, the keyword is a poor predictor of future trends, because it is too broad. The specific methods, which execute the optimization or the area where optimization is used, are better predictors.

After that irrelevance reduction, we selected 40 keywords which occurred most often in the papers. These keywords represent the main topics covered in CBM.

For the dynamic keyword change analysis, we have split the 40 selected keywords into four subsets according to the keyword area. These four areas are roughly aligned with concepts and methods for computers in biology and medicine. The following list details the keyword areas:

1. M – This subset includes medical terms as well as diseases. Fighting diseases and improving medial processes provides a powerful justification for computers in biology and medicine.
2. D – Includes both medical data and data acquisition methods. That data constitutes the input for the computer algorithms.
3. F – Includes both features and feature extraction methods. Feature extraction algorithms extract relevant information from medical data.
4. AI – Includes methods for automated decision making. Computer based decision making is important for diagnosis support.

2.3. Keyword clustering

There are two map types commonly used in bibliometric research [39]. These types are referred to as distance-based maps and graph-based maps. Distance-based maps establish the distance between two items. To be specific, the distance reflects the strength of the relation between the items [40]. We used the VOSviewer software to carry out distance based keyword clustering [41–43]. The distance between the selected keywords is established by counting the number of papers that contain both keywords. A large number of co-occurrences indicates a short distance between the selected keywords. That distance is reflected in the co-occurrence map which is used to visualize the clusters [44,45].

2.4. Keyword change analysis

Keyword change analysis was conducted with three different tools. In the first step, the bibliographics [46] package for R [47] was used to extract the keywords together with the year of publication information. The R code implements an algorithm which counts how often a specific keyword appears in a year. In effect that created a quantisation where all 15707 keywords were mapped into 28 sequential time slices.

Over time, the number of publications, and hence the number of keywords fluctuates. For CBM, there is a clear trend towards more publications in recent years, as shown in Fig. 2. That creates a problem if we take the number of keywords, within a year, as a measure of keyword
importance. To be specific, it makes a difference if a keyword appears once in 24 publications or once in 311 publications. Clearly, the keyword appearing once in 311 publications is more important than the keyword that appears once in 311 publications.

To address the fluctuation problem, we have normalized the number of appearances for a specific keyword within a year \( SY; K \) by the number of appearances for all the keywords within a year \( NY \). The following equation defines the normalization:

\[
\text{Normalized appearance in } \% = \frac{SY; K}{NY} \times 100
\]  

where \( Y \) is the year and \( K \) is the keyword. For example, the keyword SVM appeared in 2015 48 times and the number of appearances for all keywords was 1357. Plugging these values into Equation (1):

\[
\text{Normalized appearance in } \% = \frac{48}{1357} \times 100 = 3.537 \%
\]

The results were saved as CBM_all.csv. That file was loaded into the Matlab environment. The Matlab script extracted a time-line for each topic. The function result = timeline (...) plugs the gaps within a time-line. To be specific, when there was no keyword detected, within a specific year, the function inserted a 0 in the time-line. A separate csv file is saved, with the time-line data for each of the selected keywords. For example, Table 2 shows the SVM.csv file content. Column 1 lists the 28 time-slices, labelled as the year of publication. Column 2 states the normalized appearance in %. In 1990, SVM did not appear in any CBM paper, therefore a 0 was inserted. The time slice 2015 (highlighted) links the example, shown in Equation (2), with the content of the SVM.csv file, shown in Table 2.

In a final step, 40 csv files, with the analysis results for the selected keywords, were loaded into the text processing system LaTeX [48]. The data is displayed with the pgfplots package [49]. The package was also used to generate a trend-line for each graph. Mathematically, that trend-line is a linear approximation [50], which takes the form:

\[
\text{trend line}(Y) = b + mY
\]

where \( Y \) is the year, \( b \) is an offset and \( m \) is the trend-line gradient. The following section presents a plot of the normalized yearly appearance for each selected keyword as well as the static analysis results.

3. Results

The result presentation starts by listing the selected keywords. That is done in Column 1 of Table 3. This column states the acronym and, if not done so in the text before, the acronym is introduced. Column 2 provides the keyword area. We found: 7 keywords for Medical (M), 10 keywords for Data (D), 17 keywords for Feature (F) and 6 keywords for Artificial Intelligence (AI). The header C in Column 3 indicates the cluster

![Normalized appearance in % of keywords in the area of Medical (M).](image-url)
The occurrence, reported in Column 4, indicates how often the keyword appeared in the 2946 papers. In total the 40 selected keywords appeared 1117 times. Column 5 reports the trend-line gradient, which is a result of the topic change analysis. The sum of the 40 gradient values is $78.93 \times 10^{-2}$. Compared to the individual $m$ values, this is a large positive number. Being large and positive indicates that the keyword diversity increases over time.

The network graph, shown in Fig. 3, visualizes the distance based clustering result. The node colour was assigned as follows: Blue to cluster 3, light blue to cluster 6, green to cluster 2, red to cluster 1, purple to cluster 4. We discuss the clustering result in Section 4.

Fig. 4 shows the 7 graphs for the keyword change analysis in the area M, for diseases and medical terms. The graphs are ordered in terms of their trend-line gradient $m$. A positive gradient shows that the keyword gained importance in the CBM journal over the observed period. Conversely, a negative gradient indicates that the topic lost importance. In the upper left corner of Fig. 4, we start with ATRIAL FIBRILLATION, because its trend-line has the steepest ascent, i.e. the highest $m$ value. The subsequent plots have a decreasing trend-line gradient value. In effect that orders the keywords in terms of their importance for CBM. The arithmetic mean of all trend-line gradients for the area M is $3.47 \times 10^{-2}$. That means, the keyword change dynamics in CBM are such that the keywords for get more important.

Fig. 5 shows the 10 graphs for the keyword change analysis in the area D, for data and data acquisition methods. The graphs are ordered in terms of the trend-line gradient $m$. We start in the upper left corner with plotting normalized appearance of ECG over the publication years, because this keyword has the highest trend-line gradient in the area., shown in the lower left corner, has the trend-line with the steepest decent. The arithmetic mean of all trend-line gradients for that area is...
2.52·10^{-2}.

Figs. 6 and 7 show the 17 graphs for the keyword change analysis in the area F, for features and feature extraction methods. The graphs are ordered in terms of the trend-line gradient \( m \). We start from the steepest ascent, for SEGMENTATION, and progress to the steepest decent, for PHARMACOKINETICS. The arithmetic mean of all trend-line gradients for that area is 1.21·10^{-2}.

Fig. 8 shows the 6 result graphs for the keyword change analysis in the area AI, for soft computing methods and decision support. The graphs are ordered in terms of the trend-line gradient \( m \). We start from the steepest ascent, for SVM, and progress to the steepest decent, for SVM. The arithmetic mean of all trend-line gradients for that area is 1.50·10^{-2}.

4. Discussion

We start this section by discussing the distance based clustering results. The results were presented in Column 3 of Table 3 and the diagram in Fig. 3 visualizes the clusters. The following paragraphs highlight the main properties of each cluster.

Cluster 1 is centred on IMAGE PROCESSING [51–54]. These images can come from MRI [55,56], ULTRASOUND [57,58] or [59,60]. As such, MYOCARDIAL INFARCTION can be diagnosed using ultrasound images [61,62], therefore this topic is included in cluster 1.

Cluster 2 is centred on ECG [63] signals. HRV is extracted from ECG signals [64] and as such it is a good predictor for ANS [65,66]. The cluster reveals that FRACTAL DIMENSION [67,68], PCA [69,70], SPECTRAL ANALYSIS [71,72], and WAVELET TRANSFORM [73–75] are used to extract features from these biomedical signals. These features can be used to detect heart diseases, such as ATRIAL FIBRILLATION [76,77].

Cluster 3 is centred on the AI methods, such as SVM [78,79], ANN [80], GENETIC ALGORITHM [81], and FUZZY LOGIC [82]. These methods are often used in CANCER [83,84] diagnosis support systems. They can incorporate GENE EXPRESSION [85,86] and PHARMACOKINETICS [87,88] methods for gene analysis. These systems can also incorporate DWT, usually for feature extraction from images and signals [89–93].

Cluster 4 centres on BREAST CANCER and MAMMOGRAPHY [94, 95]. For automated diagnosis support NN can be used [96]. GENE EXPRESSION DATA can be analysed with CLUSTERING [97,98].

Cluster 5 is centred on ATHEROSCLEROSIS and the feature extraction methods associated with that disease. To be specific, the feature extraction methods are BLOOD FLOW [99,100], FLUID DYNAMICS [101], FEM [102], WALL SHEAR STRESS [103].

Cluster 6 is centred EEG signals [104,105]. These signals can be used to diagnose EPILEPSY [106,107].

The yearly keyword change result graphs, shown in Figs. 4–8, reveal the dynamic topic changes in CBM. We found that the topic area of AI moves fast. The graphs and trend-lines, shown in Fig. 9, document that ANN was superseded by SVM. The trend-line for ANN has the steepest decent for all the analysed keywords. In contrast, the trend-line for SVM has the steepest ascent for all analysed keywords. Both methods are in the same area, namely AI, hence they are in direct competition. According to the trend-lines, SVM superseded ANN around 2006. That paradigm shift indicates that AI is a very active area with research being conducted far
beyond medicine and biology. Therefore, we expect further topic changes in that area. For example, deep learning has the potential to supersede SVM as the predominant method [108–110].

The yearly keyword distribution reveals that GENETIC ALGORITHM is a new topic in CBM. It started trending in 2005 and it has the second largest trend-line gradient of all AI methods [111,112]. SEGMENTATION [113] has the highest trend-line gradient within F. That feature extraction method is important for computer aided diagnosis of retinopathy [114–116] and cancer [117–119]. ATRIAL FIBRILLATION is associated with significant morbidity and mortality [120]. That cardiac arrhythmia is a risk factor for ischemic stroke [121]. Therefore, it is not supersizing that ATRIAL FIBRILLATION has the highest trend-line gradient of group M. Similarly, ECG has the highest trend-line gradient for group D. That is hardly supersizing, because that physiological signal is used to diagnose a wide range of heart diseases [122]. One particular interest for ECG processing is QRS detection, which is used for HRV analysis [123–125]. Furthermore, ECG is the signal of choice for arrhythmia detection [126–128]. That versatility explains the growing levels of interest in ECG.

The declining trend-lines for IMAGE ANALYSIS and MEDICAL IMAGING indicate that these methods don’t find their way into CBM any longer. Similarly, MRI shows also a declining trend-line. IMAGE ANALYSIS and MEDICAL IMAGING are used to extract features from medical images, such as MRI [129–131]. Outside the scope of CBM, all three topics are important, because these methods have a significant role to play in the diagnosis of soft tissue cancer [132–134] and other dangerous diseases [135,136]. Furthermore, the trend for imaging data goes towards more images and higher resolution [137]. Computer assisted feature extraction can be used to avoid data overloading of the reading radiographer [138]. The mediocre success of CBM to attract publications in these areas can be attributed to the competition from specialized journals. We suspect that these journals take the majority of papers about IMAGE ANALYSIS, MEDICAL IMAGING and MRI. Diseases create the need for computer based problem solutions. Therefore, we were surprised that diabetes [139] did not feature in the selected keywords; neither did diseases related to ageing, like Alzheimer and dementia. We can only speculate on the reasons why CBM could not attract more papers on these topics. There are other journals that may take these papers. Real-time is another topic that did not feature in the 40 selected keywords, despite the fact that real-time is mentioned twice in the journal aims. One reason for that lack of real-time related papers might be that modern computing machinery is fast enough for most applications [140]. Furthermore, this topic is very specialized, hard real-time is only needed for control applications, such as smart prostheses and brain stimulation [141]. Therefore, we suspect that papers, which address real time issues, go to journals with a narrow focus on that area. Another point is that none of the selected keywords is related to safety and reliability of computer systems for biology and medicine. This is a big and growing problem, because most scientists are working on creating ever more complex problem solutions and only a minority reflects on how to ensure safety and reliability [142,143]. To be specific, CBM does not publish research work that indicates whether these complex systems
have dangerous side effects. The lack of research in that area has far reaching consequences, because safety and reliability concerns are a big hurdle when it comes to large scale deployment of computer executed algorithms for medicine and biology [144].

Having discussed our analysis results, we move on to present some insights into the processes which influence topic change. There are technical aspects which cause one topic to be superseded by another. For example, ANN is in decline because SVM delivers better classification results [145]. However, there are topic changes which cannot be explained by technical aspects alone. For example, the normalized appearance graph for IMAGE REGISTRATION, shown in Fig. 7, peaked in 1999. Subsequently, the keyword did not appear in any paper for 6 consecutive years before it picked up again. Similar gaps exist for BLOOD FLOW and CT. We suspect that these gaps indicate a problem in the CBM management. Fig. 2 shows that CBM did not attract many papers during the period from 2000 to 2004. That means the gaps in the normalized appearance of keywords plots coincide with a period of low article output. Since 2005, the number of articles published in CBM increased significantly. A lack of focus and effort from the editorial board is likely to have caused that downturn.

The editorial board and the editor in chief in particular play a crucial role in the journal success. CBM has a broad scope, hence good quality control is needed to attract and publish outstanding papers from a wide range of biomedical research fields. Reviews with a high standard should be provided to authors as fast as possible. That will encourage authors to submit their good papers and this in turn will improve the reputation of the journal. Another way to increase the topic diversity, while maintaining quality control, is to incorporate special issues. The guest editors need to be pioneers in the special issue areas, in order to attract good quality submissions. For example, Fig. 2 shows that the year 2007 had 197 articles, considerably higher than all the previous years and the following 5 years. Closer inspection shows that there were 12 issues in 2007, the same number as in 2006 and 2008. However, there were 3 special issues during 2007. In 2008 there was no special issue published and in 2006 there were 2 special issues. The special issue on medical ontologies was published in July and August 2006, i.e. as issues 7 and 8. That means, there were only 11 content issues in 2006.

This study depends on the web of science database. There are other sources, such as EMBASE and MEDLINE available. However, the alternative sources provide the data in a different format. Therefore, using them was impracticable. Furthermore, the CBM publications are listed in the web of science database – that was sufficient for the current study. The main limitation of this study is the subjective selection of 40
keywords. Using more keywords might reveal further trends and topic changes. Another limitation is that we have used linear models to predict future topics from past observations. We recognise that computers and related topics are the engine of change. That change is rather predictable and, according to Moore's law, it is almost linear [1-16]. However, a new medical data acquisition method or a new widespread disease may become a game changer. Unfortunately, these game changers are hard to predict, especially with linear models.

5. Conclusion

In this study, we documented and predicted topic changes in CBM. On a technical level, that was done by analysing author supplied keywords in all CBM publications from 1990 to 2017. The analysis was structured into two distinct parts. The first part was cluster analysis and the second part focused on tracking the keywords over time. The cluster analysis revealed the depth and breadth of topics covered in CBM. Tracking keywords over time resulted in 40 graphs that document the normalized yearly keyword appearance in percent. A large value of the accumulated gradients provides strong evidence for the diversity of CBM topics.

Our bibliometric analysis can help the CBM readership and the wider research community to plan their research. We found that SVM, EEG and IMAGE PROCESSING are very well established topics in. In the discussion section, we put forward some research gaps and highlight deep learning as an up and coming decision making algorithm. We recognise that computers in biology and medicine will continue to be an engine for progress and indeed change. That change will be gradual and linear, similar to the gradual demise of ANN and the steady rise of SVM. We do not anticipate any sudden and unexpected events from using mature computing machinery. Such sudden and unexpected events can occur as a reaction to new wide spread diseases and, to a lesser extent, in reaction to new data acquisition methods. The outcomes of our analysis can help the CBM editors to maintain the journal as a relevant forum to exchange ideas on the latest trends for computing machinery in biology and medicine. We recommend embracing new ideas and not shying away from complex methods, because they are the agents of change. Fast and sound reviews will improve the reputation with authors. Frequent guest editions can also help to explore new topics or vitalize dormant areas.

References


