Scientific Visualization, 2024, volume 16, number 5, pages 1 - 21, DOI: 10.26583/sv.16.5.01

Visualizing the Impact of Machine Learning on Cardiovascular Disease Prediction: A Comprehensive Analysis of Research Trends

Jeena Joseph[1,A,B], K Kartheeban[2,C]

[A] Department of Computer Applications, Kalasalingam Academy of Research and Education, Krishnankoil, Tamilnadu, India
[B] Department of Computer Applications, Marian College Kuttikkanam Autonomous, Idukki, Kerala, India
[C] Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Tamilnadu, India

[1] ORCID: 0000-0003-4070-5868, jeenajoseph@mariancollege.org
[2] ORCID: 0000-0002-0001-3376, k.kartheeban@klu.ac.in

Abstract

Cardiovascular diseases (CVDs) continue to have a negative impact on global health, which highlights the need for accurate and efficient prediction methods. Machine learning (ML) techniques as tools for forecasting CVD has recently showed potential. This paper presents a comprehensive analysis of research trends in the field, focusing on visualizing the impact of ML in cardiovascular disease prediction. We used data visualization techniques to identify patterns and trends in an extensive database of scholarly publications on this subject that were published in Scopus between 1991 and 2023. The analysis reveals a substantial growth in research output, demonstrating the growing demand for ML-based CVD prediction. It reveals essential stakeholders and potential collaborators while highlighting the institutions and authors who have contributed most to this domain. The study also identifies high-impact journals that have published significant research in this domain, facilitating researchers in selecting appropriate outlets for dissemination. The study helps researchers identify the most critical areas for further research and fosters cooperation among subject-matter experts by offering insightful information about machine learning-based cardiovascular disease prediction development. The data is analyzed using the tools VOSviewer and Biblioshiny.

Keywords: Bibliometric analysis, Cardiovascular disease, Heart disease prediction, Machine Learning, Biblioshiny, VOSviewer.

1. Introduction

Heart disease is one of the most prevalent causes of death worldwide, posing a major problem for public health systems [1]. The World Health Organization (WHO) predicts that heart failure and stroke will account for the bulk of the 23.6 million fatalities from cardiovascular disease (CVD) worldwide in 2030 [2,3]. Effective prevention and prompt intervention depend on early detection and precise cardiac disease prediction. Clinical risk factors, medical history, and diagnostic tests are the mainstays of conventional methods for diagnosing heart disease. Fortunately, cardiovascular disease prediction has been a paradigm shift with the introduction of machine learning techniques and the growing amount of digital health data [4].

With the potential to increase the precision and efficacy of cardiac disease prediction, machine learning, a branch of artificial intelligence, has become a valuable tool in the healthcare

industry [5,6]. Machine learning algorithms can spot complicated patterns and associations that are difficult to spot using conventional approaches by utilizing the enormous quantity of data created by electronic health records, wearable technology, and medical imaging [7]. Some supervised algorithms, including Naive Bayes, Random Forest (RF), Decision Trees (DT), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN), help predict cardiovascular disease (CVD) [8].

Bibliometric analysis is a recognized statistical tool to assess and visually depict the current state and research patterns within particular disciplines, which is widely utilized across many fields [9-14]. It provides a fair assessment of scientific contributions by carefully reviewing pertinent works from authors, organizations, nations, journals, citations, and keywords. Additionally, it aids in locating current trends and research priorities, which deepens understanding of present trends [15-16].

Bibliometric networks can be visualized using the robust and user-friendly software tool VOSviewer [17- 20]. Users can create maps and visual representations of bibliometric data through co-citation analysis, bibliographic coupling, and co-authorship analysis [15,17,18]. Researchers can find prospective collaborations and new trends in their fields using these maps, an invaluable resource. Researchers can acquire new insights from them by identifying clusters of related study topics, well-known authors, and institutions. However, Biblioshiny is a fantastic tool for scholars who want to show their bibliographic data in an exciting and lively way [21]. R programming uses the “shiny” package, making it easier to create interactive web applications. Using Biblioshiny, academics, and researchers may create flexible, individualized bibliographies that make it simple to quickly explore and analyze their sources. Users who want to connect with and study their research materials will have a better experience if bibliographic data is integrated into the package.

The aim of this bibliometric analysis is to provide a thorough overview of the research on utilizing machine learning to predict cardiovascular disease. This is accomplished by carefully reviewing the pertinent literature published in this study domain. The examination covers various topics, such as research trends, bibliographic coupling, international cooperation, and the networking of scholars in this area. For this study, information was gathered using data from the Scopus database, which covers the years 1991 through 2023. Advanced bibliometric visualization tools like VOSviewer and Biblioshiny were used to create a precise and up-to-date representation of the prevalent research practices. The potential research objectives of this bibliometric study are:

- To Map the Evolution of Research on cardiovascular disease prediction using machine learning by analyzing publication trends from 1991 to 2023. This includes identifying key periods of increased research activity and growth in the field.
- To Identify Key Contributors in the research domain, including prolific authors, leading journals, and institutions that have significantly contributed to advancing the study of cardiovascular disease prediction using machine learning.
- To Analyze Publication and Citation Trends to understand the impact and reach of research conducted in this area. This involves examining the number of publications over time and the citation patterns to gauge the influence of specific works.
- To Visualize the Research Landscape through the use of bibliometric tools and techniques, such as Biblioshiny and VOSviewer, to create visual representations of publication counts, author productivity, journal contributions, and institutional involvements.
- To Examine Collaborative Networks among authors, countries, and institutions to highlight the collaborative nature of research efforts in this domain and to identify the most influential networks contributing to the advancement of ML in cardiovascular disease prediction.
- To Investigate Key Themes and Topics within the research through keyword and thematic analysis, utilizing techniques like keyword co-occurrence networks and thematic mapping to uncover the core focuses and evolving interests in the field.

- To Assess the Technological Advances and methodological approaches employed in machine learning for cardiovascular disease prediction, detailing the use of specific algorithms, models, and computational techniques that have been prominently featured in research publications.

2. Literature Review

Researchers have been concentrating on creating various predictive models employing machine learning, artificial intelligence, and data mining approaches in recent years to help with the early identification and prediction of cardiac disease. Several research studies have shown the successful utilization of machine learning (ML) models in identifying heart diseases [22]. Several researchers have suggested ML algorithms to improve the precision of disease prediction [23]. A substantial amount of study has been devoted to carefully examining the presence of missing information within the dataset, which is a vital component of the data preprocessing stage, to improve the correctness of the results. To improve the accuracy of their results, Gupta et al. [24] replaced missing values in the Cleveland dataset using Pearson correlation coefficients and several Machine Learning classifiers. Mohan et al. [25] created a combination of a random forest (RF) and a linear model, known as a hybrid RF. This hybrid approach enhanced the accuracy of predicting heart disease by 297 instances and improved the performance on 13 different features from the Cleveland dataset. Furthermore, the method employed for selecting features significantly enhances the model's precision. In the study by Shah et al. [26], probabilistic principal component analysis (PCA) was employed to choose features. Another research conducted by R. Perumal et al. [27] implemented the Cleveland dataset to develop logistic regression and support vector machine models, achieving comparable accuracy rates of 87% and 85% accordingly.

The UCI Heart Disease Dataset, which the public can access via the UCI Machine Learning Repository, is a well-known dataset in this research area and is frequently utilized [28]. Similar to this, the Statlog dataset is frequently used [29]. ML models seek to improve accuracy and lower processing costs in the clinical identification of diseases. For instance, Verma et al.'s [30] hybrid model for predicting cardiovascular disease integrated K-nearest neighbor and multi-layer perceptron (MLP) classifiers with particle swarm optimization (PSO). This hybrid model achieved an accuracy of 90.28%.

Siddhartha [31] created a comprehensive dataset in 2020 by combining five well-known heart disease datasets: Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog. The new dataset incorporated the common features found in these five datasets. Alalawi et al. [32] utilized various classification algorithms to analyze the dataset, including Support Vector Machine, Decision Tree Classifier, Logistic Regression, KNN, and Naïve Bayes. In a study by Yilmaz et al. [33], the classification outcomes of the Logistic Regression, Random Forest, and Support Vector Machine approaches were assessed. Ivan Miguel Pires et al. (2021) used tests with various classifiers in their study, including the combination nomenclature (CN2) rule inducer, SVM, KNN, DT, and neural networks. Through the use of 5-fold, 10-fold, and 20-fold cross-validation approaches, they evaluated these classifiers' performance. At 20-fold, 10-fold, and 5-fold cross-validation, the DT classifier had the highest accuracy score of 87.69%, followed by SVM and SGD with 87.69% and 87.69%, respectively [34].

Niloy Biswas et al. employed a range of techniques in 2023 to choose important attributes and identify the ones crucial for predicting cardiovascular disease. They employed these selected features with six distinct machine-learning algorithms. Each algorithm produced a separate score based on the chosen features. Notably, the performance of SVM and LR algorithms stood out as more significant compared to the other algorithms [8]. According to a study conducted by Sivakannan Subramani et. al (2023), they propose that using a technique called stacking makes it possible to harness the advantages of different types of models to achieve better predictive accuracy. The suggested stacking method enhances prediction performance and boosts resilience and usefulness, particularly for individuals with a high sus-

ceptibility to cardiovascular disease [35]. Md Manjurul Ahsan and Zahed Siddique explore the utilization of machine learning (ML) in diagnosing heart disease from imbalanced data, highlighting the need for real-world applications that incorporate interpretable ML for reliable predictions. Their review of 49 studies reveals a predominance of Deep Learning and SMOTE for handling imbalanced datasets, with a growing interest in GAN-based models for synthetic data generation despite their computational demands. This comprehensive analysis, based on a meticulous search in the Scopus database, underscores the potential and existing challenges in ML-driven heart disease diagnosis, suggesting directions for future research [36]. Another study explores the intersection of machine learning (ML) and the Internet of Things (IoT) in heart disease diagnosis through heart sounds, analyzing studies from January 2010 to July 2021 across six databases. Of the 4,372 papers screened, 58 were thoroughly reviewed, revealing a significant emphasis on using wearable sensors and digital stethoscopes for heart rate monitoring, and on the application of ML algorithms in medical care. The analysis identified a growing trend in the use of intelligent services for predicting cardiovascular disorders in 22.41% of these studies [37].

3. Materials and Methods

The scientific papers for the study were obtained from the primary collection of the Scopus database. On May 31, 2023, a search was conducted in the database using specific terms such as “heart disease”, cardiovascular disease, heart attack, prediction, and machine learning. The search was refined to include only journal articles and conference papers published between 1991 and 2023. The data collected from the search were saved as CSV files containing complete records and references cited in the papers. To analyze the bibliometric aspects of the data, VOSviewer version 1.6.19 and the Bibloshiny software is utilized. The key aspects of this investigation are summarized in Table 1.

Table 1. Essential aspects of the investigation.

Description	**Results**
Search Query	(TITLE-ABS-KEY ("heart disease") OR TITLE-ABS-KEY ("cardiovascular disease") OR TITLE-ABS-KEY ("heart attack") AND TITLE-ABS-KEY (prediction) AND TITLE-ABS-KEY ("machine learning")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"cp"))
Main information about data	
Timespan	1991:2023
Sources (Journals, Books, etc)	1150
Documents	2564
Annual Growth Rate %	19.19
Document Average Age	2.39
Average citations per doc	11.88
References	76033
Document Contents	
Keywords Plus (ID)	10308
Author's Keywords (DE)	4342
Authors	
Authors	11470
Authors of single-authored docs	73
Authors Collaboration	
Single-authored docs	74
Co-Authors per Doc	5.67
International co-authorships %	21.68
Document Types	
Article	1578
conference paper	986

4. Results

The Scopus database search yielded a collection of 2564 records. These records encompass journal articles and conference papers that have been published between the years 1991 and 2023. The publication rate experienced a significant increase in 2015 and maintained its growth until 2022. To visually depict the relationship between the publication count and the corresponding year, the tool Biblioshiny was employed, as illustrated in Figure 1.

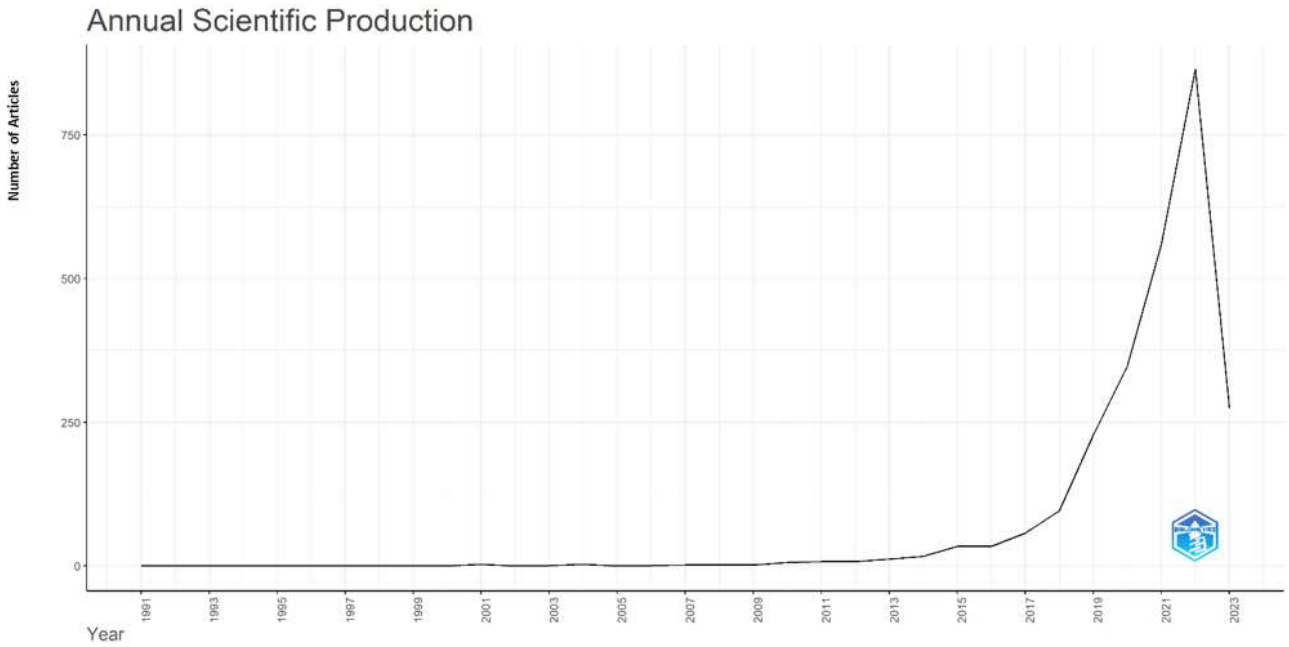

Figure 1. *The number of publications from 1991 to 2023 represented using the tool Biblioshiny. Year of publication is indicated on the x-axis, while the number of articles is indicated on the y-axis.*

4.1. Most Significant Authors

A total of 11,470 authors conducted research on predicting cardiovascular disease using machine learning. The quantity of articles published serves as a measure of their research output. Two authors, Liu Y and Wang Y, have the highest number of publications, with 25 articles credited to both of them. Liu Y follows closely with 24 articles. Table 2 displays the publication counts of the most notable authors who have published more than fifteen articles over time. These authors have established a long-standing presence in their respective fields, granting them a position of influence. Figure 2 provides a visual representation of the top ten authors with the highest productivity from 1991 to 2023. The number of articles written by each author within a specific time period was utilized to determine their level of productivity. Figure 3 presents the production of top authors over a period of time. The size of each circle corresponds to the number of articles published by the author in a given year. Additionally, the shade of each circle in Figure 2 and 3 represents the number of citations received per year, with darker circles indicating a higher citation count.

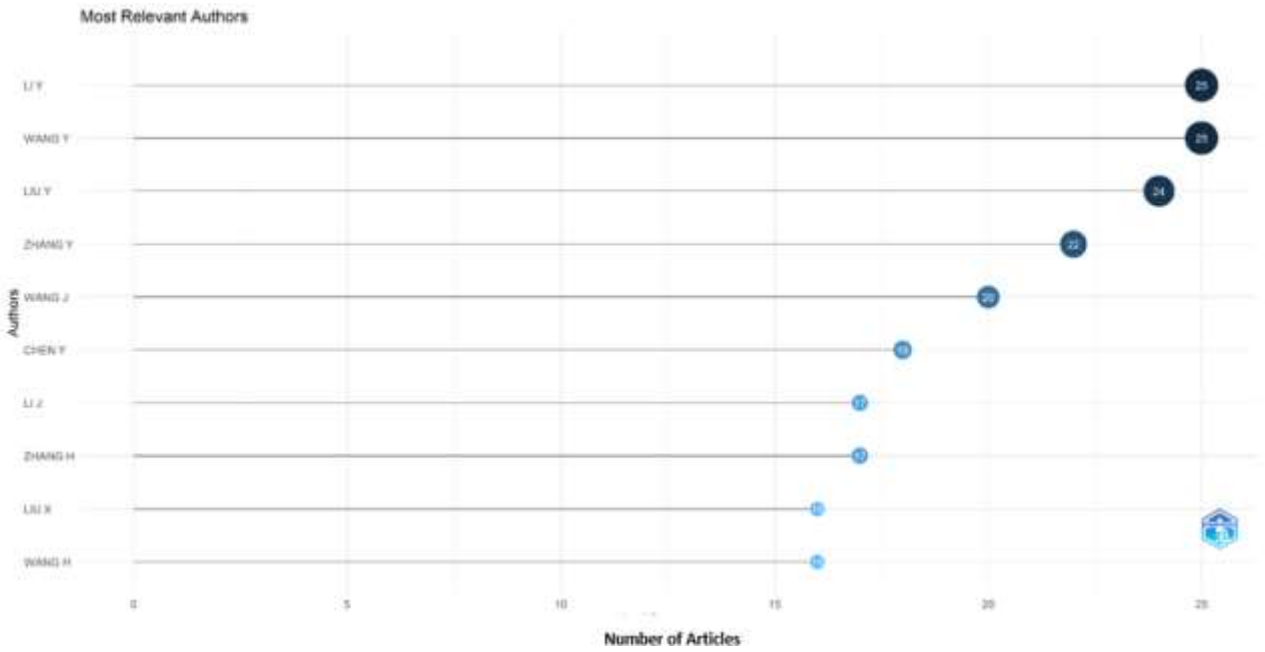

Figure 2. Most Relevant Authors. *Name of the author appears on the x-axis, while the number of documents appears on the y-axis.*

Table 2. Authors having more than fifteen articles

Authors	Articles
Li Y	25
Wang Y	25
Liu Y	24
Zhang Y	22
Wang J	20
Chen Y	18
Li J	17
Zhang H	17
Liu X	16
Wang H	16

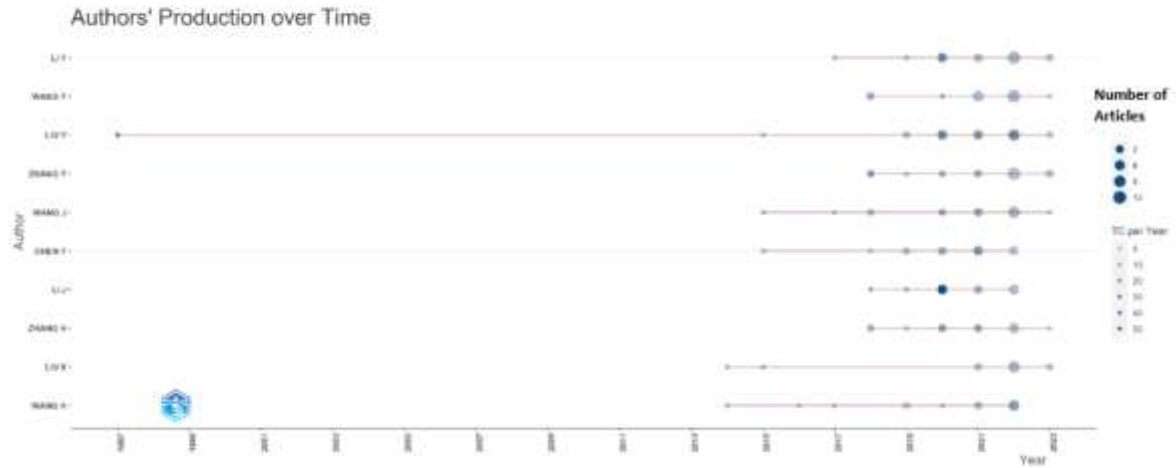

Figure 3. Top authors' production over time. *The year of publication is shown on the x-axis, while the author's name is mentioned on the y-axis.*

4.2. Most relevant sources and affiliations

A total of 2564 publications were collected from 1150 different journal sources. Among the journals analyzed, Lecture notes in networks and systems stood out as the most productive, containing a maximum of 56 articles. Following closely was Plos One, with 37 publications. Figure 4 displays the top 10 journals that published the highest number of papers in the field of Cardiovascular disease prediction using machine learning research. Moreover, the color intensity of each circle signifies the annual citation volume, where darker shades correspond to a greater number of citations.

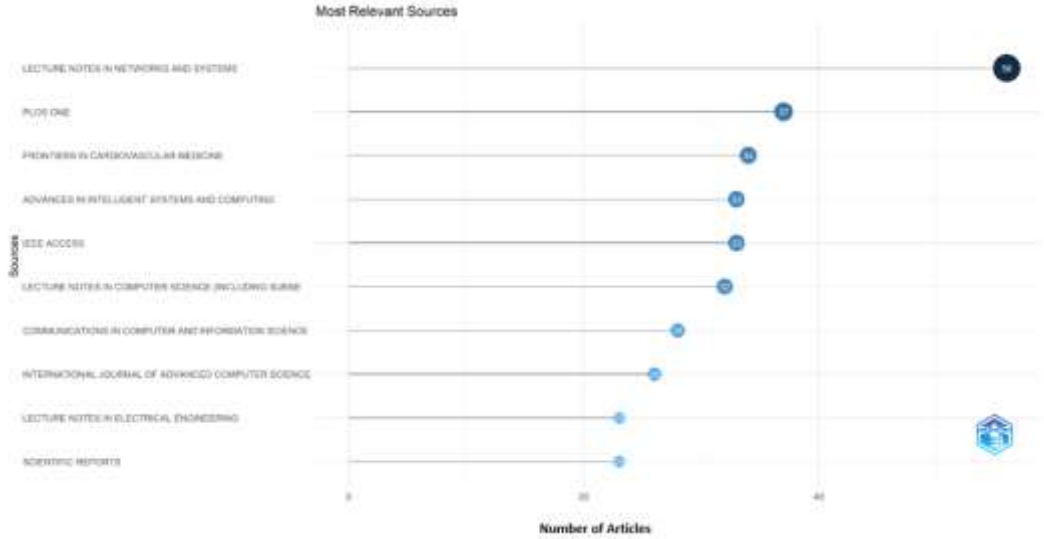

Figure 4. The top 10 relevant sources in terms of number of publications. *The number of articless is displayed on the x-axis, while the sources are depicted on the y-axis.*

Figure 5 illustrates the primary institutions involved in conducting research on predicting Cardiovascular disease using machine learning. Harvard Medical School emerges as the leading institution with a maximum of 60 research publications, closely followed by the University of Oxford with 53 publications. Notable research in this field has also been conducted at Chungbuk National University, Capital Medical University, and Stanford University, which are among the affiliated institutions with significant contributions. Furthermore, the darkness of each circle corresponds to the annual citation count, where darker circles signify a greater number of citations.

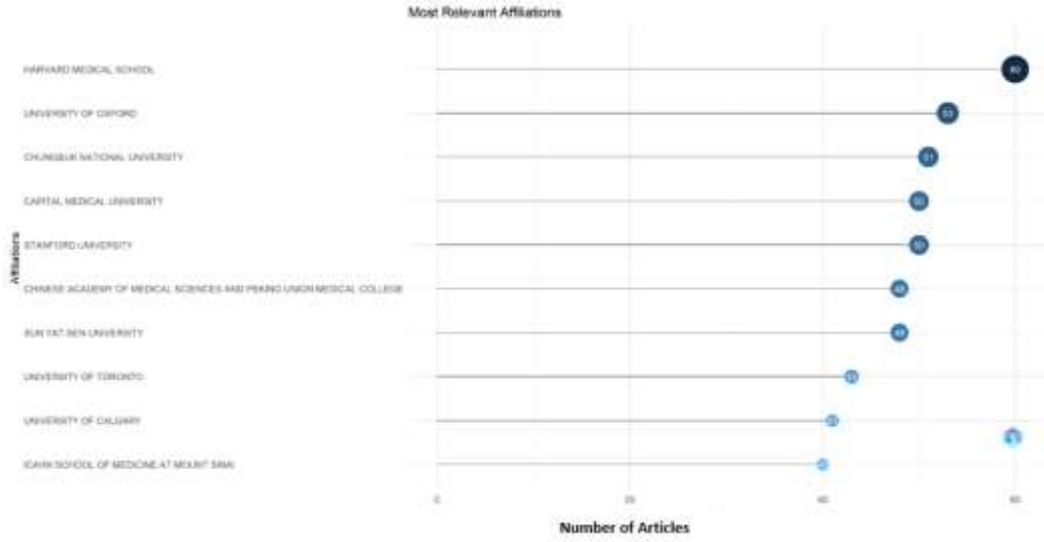

Figure 5. Most relevant affiliations in terms of the number of publications. *The x-axis lists the total number of publications, while the y-axis lists the pertinent affiliations.*

4.3. Three Field Plot on Cardiovascular Disease Prediction using Machine Learning Research

Sankey diagrams are widely employed to illustrate the movement of materials in different networks and methods. They utilize measurable attributes to depict the flow, connections, and transitions. Sankey charts are weighted graphs that ensure the preservation of flow, where the inflow weights at each node precisely match the outflow effects. These diagrams enable the visualization of processes and facilitate the exploration of relationships [38]. In Biblioshiny, the three-field plot is utilized to visually represent the connections between information sources, countries, affiliations, key phrases, prominent authors, citations, author keywords, and more [39]. Important elements are depicted as colored rectangles, with the height of the rectangle representing the association between components such as countries, organizations, sources, significant authors, keywords, etc. The width of the rectangle indicates the complexity of interactions between different components, with wider rectangles signifying more intricate connections [39].

Figure 6 presents an illustration that explores the interconnections among keywords, authors, and sources in scientific literature related to the prediction of cardiovascular disease using machine learning. The study aimed to identify frequently employed keywords in various publications and determine the authors and sources associated with them. Among the keywords examined, "machine learning" emerged as the most commonly used one. The analysis of the top keywords, authors, and sources revealed several prominent terms, including "machine learning," "random forest," "cardiovascular disease," "heart disease," and "prediction" and notable authors such as Chen Y, Li Y, Zhang Y, Wang Y, Sharma A, and Gupta A were found to have published their work in reputable sources like Frontiers in Cardiovascular Medicine and Computers in Biology and Medicine.

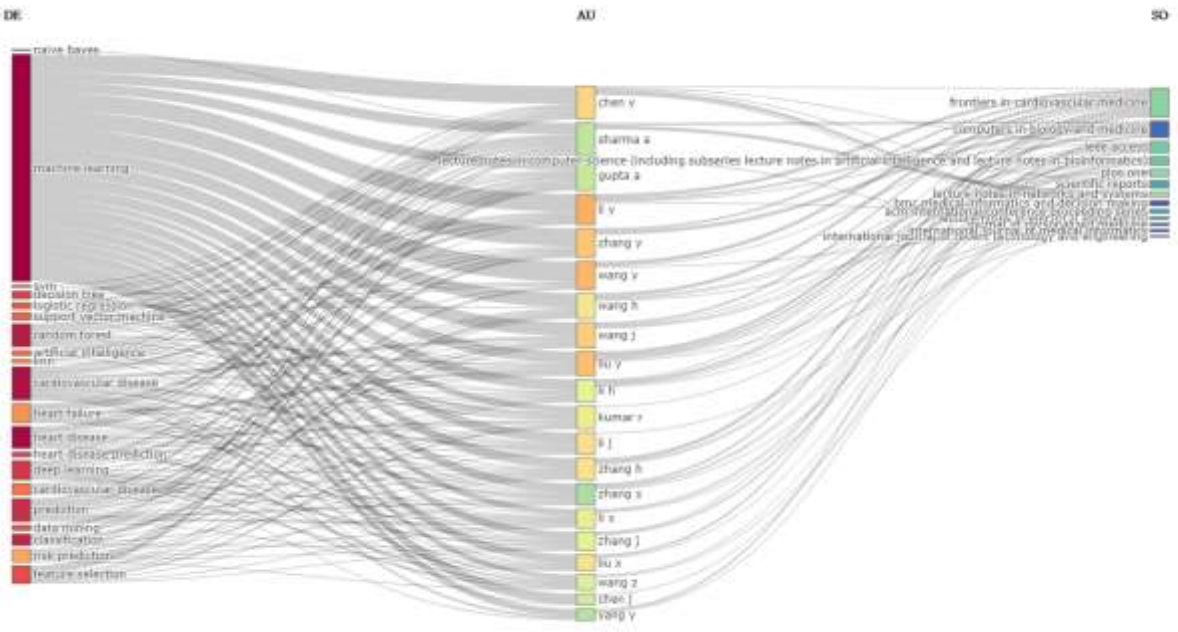

Figure 6. *Three field plot keyword (**left**), author (**middle**) and source (**right**) using Biblioshiny. (DE signifies author keyword, AU represents the author, and SO denotes the source)*

4.4.Co-Occurrence of Keywords and Content Analysis

Building a keyword co-occurrence network involves treating each keyword as a node in the network. Whenever two terms appear together, they connect those two phrases, symbolizing their association. The strength of this connection is determined by how frequently the pair of terms co-occur. Figure 7 illustrates a word network that utilizes the co-occurrence of keywords to uncover meaningful connections and research themes.

Figure 7. *Co-occurrence network generated using Biblioshiny.*

Each node in the diagram represents a keyword, and the edges connecting pairs of nodes represent the co-occurrence of those keywords. The thickness of an edge indicates how frequently the keywords appear together, while the size of a node and its label indicate the frequency of occurrence of a specific keyword. A thicker edge suggests a strong relationship between the keywords. The color of a node indicates the cluster to which the keyword belongs. The keywords and their linkages imply that each cluster is associated with a distinct research area.

The figure displays two separate clusters that Biblioshiny dynamically identifies: a red cluster and a blue cluster. The red cluster primarily encompasses keywords related to heart disease and its prediction, such as "heart disease," "predictive model," "risk prediction," "coronary artery disease," and "electronic health records." The blue cluster includes various machine learning algorithms and methods, such as "neural network," "supervised learning," "decision trees," and "support vector machines." "Machine learning" emerges as the central node in the co-occurrence network, reflecting its predominant presence within the body of literature concerning the application of computational methods to cardiological data. This node's prominence, emphasized by its size, denotes a high frequency of occurrence, signifying the term's relevance and centrality to the field. Adjacent to this pivotal node, terms such as "heart disease," "predictive model," "risk prediction," "coronary artery disease," and "electronic health records" are also highlighted, showcasing the specific focus on leveraging machine learning for diagnosing and prognosticating heart-related conditions. The clustering of these terms suggests thematic connections that warrant further investigation for a more nuanced understanding of their interrelationships.

It's important to note that the separation between the red and blue clusters is not perfect, as evidenced by the intersection at terms like "neural network," which appears in both clusters. This overlap indicates areas where research themes converge, reflecting the interdisciplinary nature of the field. The network's edges, varying in thickness, map the strength of the co-occurrences, with thicker lines illustrating a more frequent association between terms. This co-occurrence network thus not only serves as a testament to the interdisciplinary nexus between machine learning and cardiac healthcare but also aids in identifying prevailing research trends and potential avenues for future inquiry.

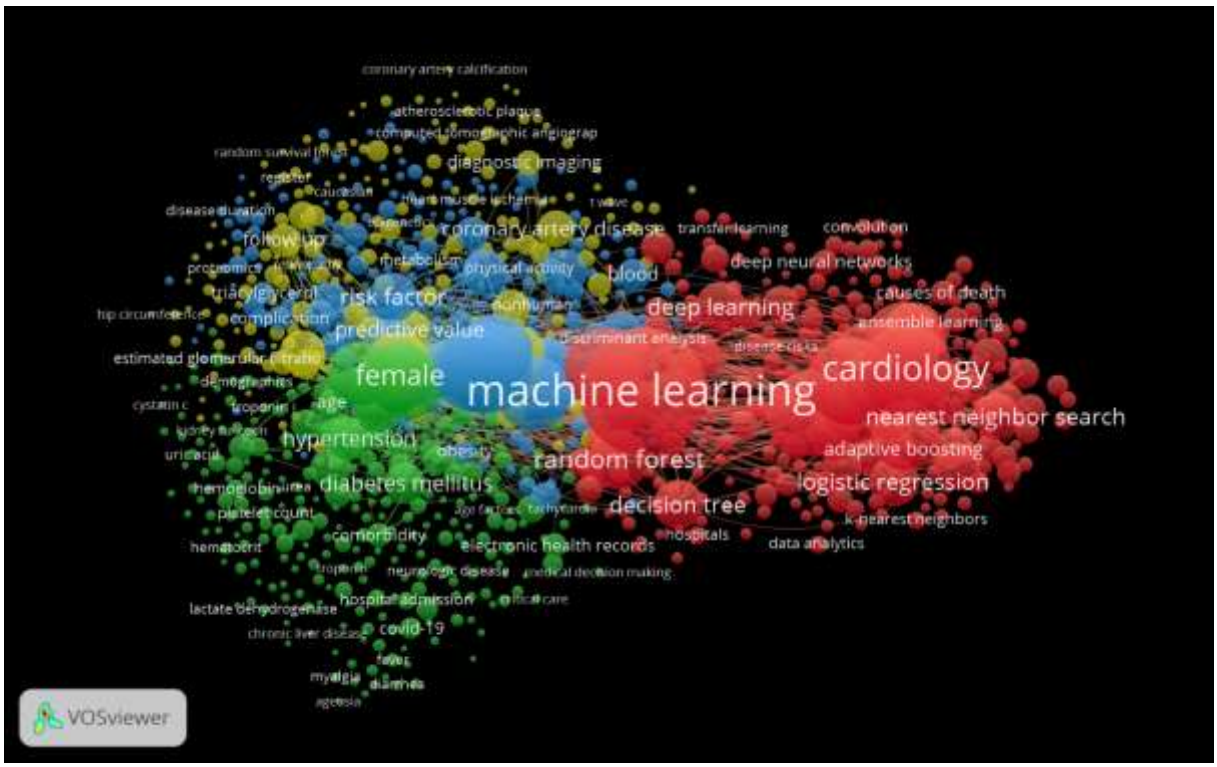

Figure 8. Network Visualization of keyword co-occurrence using VOSviewer

Figure 8 displays a visual representation where the size of nodes and font is determined by the weight assigned to a keyword. A higher weight value indicates more frequent keyword occurrences, resulting in larger nodes and fonts. The lines connecting the nodes indicate common appearances of keywords. The thickness of these lines represents the strength of co-occurrence between two keywords. Thicker lines indicate a higher frequency of co-occurrence. Figure 8, identifies four clusters. Each cluster encompasses a specific topic, and a list of keywords associated with each cluster. This elucidates the interconnectivity between terms within the domain of machine learning and its application to cardiology. The nodes, varying in size, signify the frequency of each keyword's occurrence, with larger nodes such as "machine learning," "female," and "cardiology" indicating a higher prevalence in the dataset. These keywords act as anchors within their respective clusters, which are differentiated by color and encapsulate specific thematic areas. The red cluster prominently features "cardiology," "logistic regression," and "decision tree," pointing towards a strong focus on machine learning algorithms applied to cardiovascular research. The green cluster encapsulates terms like "female," "diabetes mellitus," and "hypertension," suggesting an emphasis on demographic and clinical factors in machine learning research. Yellow nodes, with "risk factors" and "blood" as notable examples, may indicate a concentration on the identification and analysis of risk factors within medical datasets. Lastly, the blue cluster, highlighted by "deep neural networks" and "ensemble learning," points towards a technical focus on sophisticated, often multilayered, machine learning models. The clear demarcation of clusters underscores the multidisciplinary nature of the research, bridging complex algorithmic approaches with practical clinical applications. The connections between nodes, particularly the thicker lines, denote strong co-occurrences and imply a substantial cross-disciplinary dialogue. The detailed description of these clusters within the article will provide readers with a nuanced understanding of how machine learning is woven into the fabric of contemporary medical research, particularly within the realm of cardiology, reflecting the current trends and potential future directions of the field.

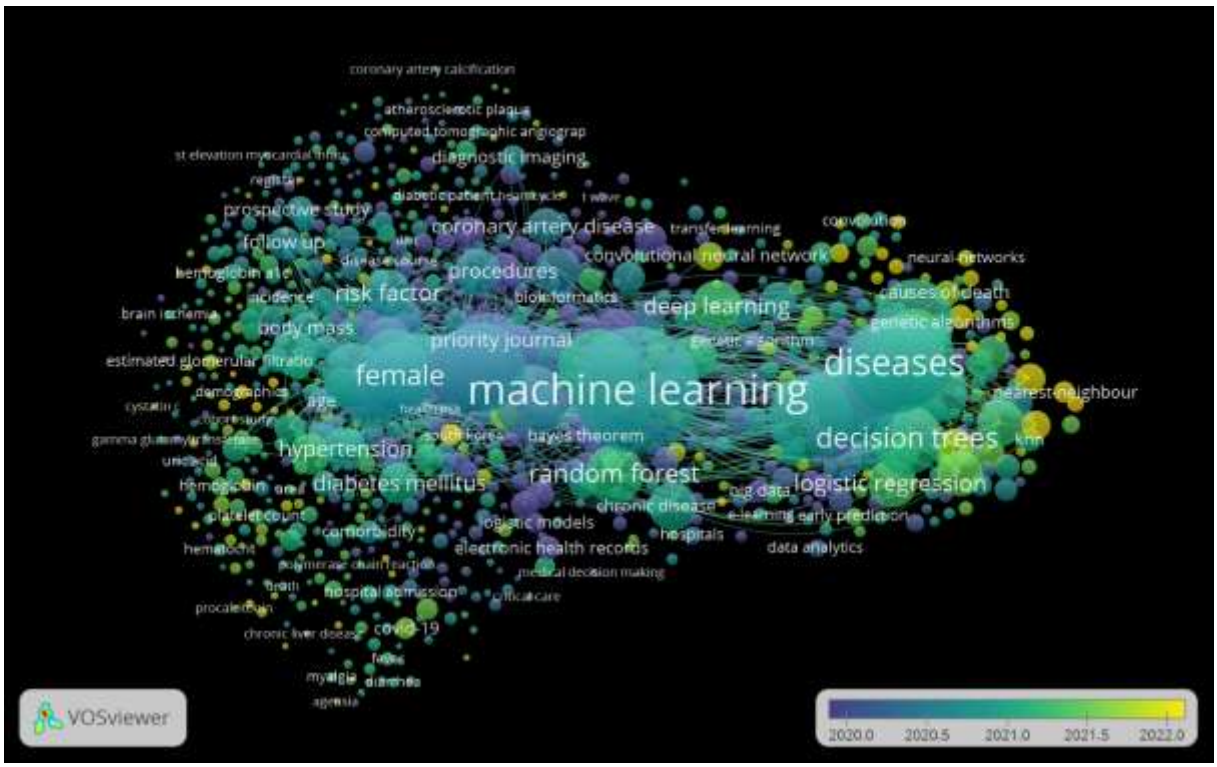

Figure 9. Overlay Visualization of keyword co-occurrence chronology using VOSviewer

VOSviewer examines the occurrence of keywords over different years, providing insights into the progression and development of research in the field of cardiovascular disease prediction using machine learning. Figure 9 displays a visual representation of this chronological view. The color of the lines connecting the keywords signifies the initial occurrence of their co-occurrence, where blue represents earlier years and yellow represents more recent years. This color-coding helps identify when specific keywords began to appear together in the literature. When terms span the entire time range, their colors might blend, indicating their persistent relevance over time. The thickness of the lines indicates the strength and frequency of co-occurrence between the two keywords. Thicker lines represent stronger and more frequent connections. By examining these visual cues, insights into the progression of research can be gained. For instance, earlier co-occurrences (shown in blue) highlight foundational concepts and methodologies like "random forest" and "logistic regression." More recent co-occurrences (shown in yellow) highlight emerging trends and advanced techniques such as "deep learning," "genetic algorithms," and the impact of "COVID-19" on cardiovascular disease prediction research.

4.5. Most Frequent Words and Word Cloud of the keywords

Figure 10 illustrates the key phrases that are commonly utilized and their respective frequencies, as identified by the biblioshiny software. The highest ten keywords were as follows: "machine learning (1859)", "diseases (1088)", "cardiology (968)", "human (922)", "forecasting (879)", "female (810)", "article (807)", "male (793)", "heart disease (782)", and "cardiovascular disease (758)".

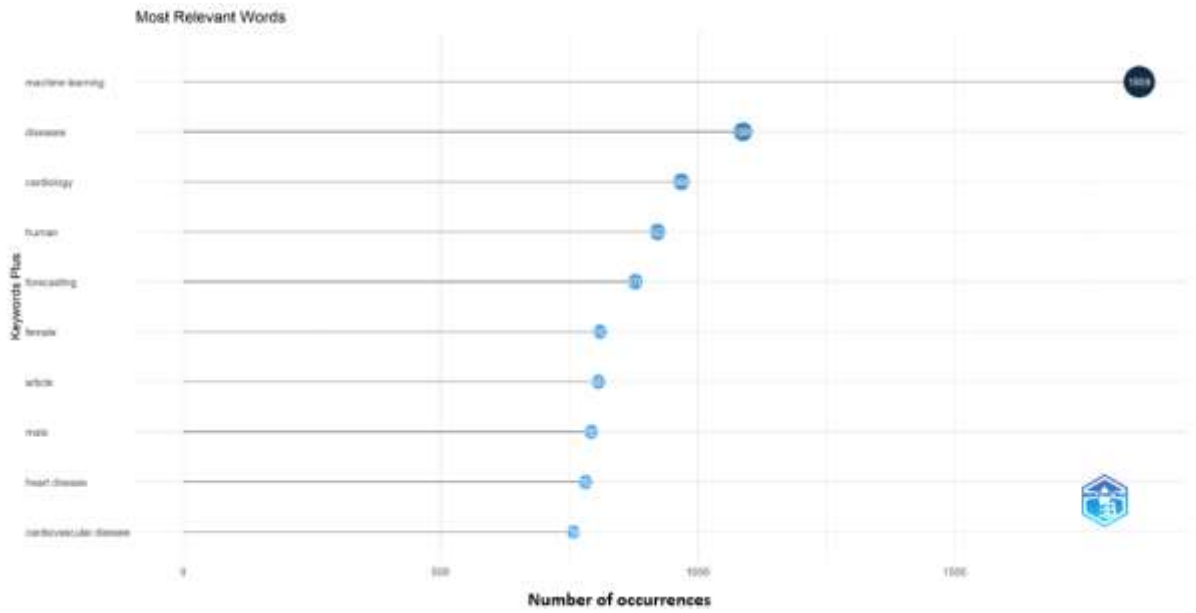

Figure 10. The most frequently used keywords. The x-axis represents the number of publications and the y-axis represents the keywords plus.

Figure 11 and Figure 12 depict the authors' keywords and keywords plus, represented in a word cloud. Authors' keywords are chosen by the researchers to directly reflect the central themes or concepts of their study, ensuring that their work can be easily found and recognized for its main contributions. In contrast, Keywords Plus are generated by the indexing algorithms of academic databases, based on the titles of cited articles. This extends the article's reach by connecting it with a wider array of related topics and themes, thereby enhancing its discoverability. The purpose of using a word cloud to analyze the articles is to examine the frequently appearing phrases, focusing on those areas for analysis. A word cloud converts text input into identifiable terms, typically short ones, where their size in the cloud indicates their relative importance. This facilitates connections between the main keywords, which are distributed among the most significant terms such as "machine learning," "cardiology," "cardiovascular disease," "heart disease," "forecasting," and so on. In elucidating the mechanism through which word clouds facilitate connections among keywords, it becomes apparent that their effectiveness is rooted in a combination of visual cues and cognitive processes. The differing font sizes employed in word clouds serve not just as an aesthetic choice but as a visual indicator of frequency, with larger sizes denoting higher frequency terms. This feature immediately highlights the most significant themes within the dataset, guiding the viewer's attention to key areas of interest. While the spatial arrangement of words—particularly their proximity and potential grouping—can vary across different word cloud generators, when applied, this arrangement subtly suggests relationships between terms that frequently appear together, offering insights into thematic clusters. Moreover, the aggregation of keywords into a single visual tableau facilitates a direct comparison of their relative prominence, prompting further inquiry into their interrelations and the broader thematic contours they outline. In academic contexts, this visualization strategy proves invaluable. By amalgamating authors' chosen keywords with the automatically generated Keywords Plus, word clouds not only spotlight the central themes of a study but also map out a wider landscape of related topics, thereby broadening the viewer's perspective. This visualization capitalizes on the human propensity to process and interpret visual information efficiently, making word clouds a powerful tool for uncovering and understanding the complex web of connections that define the thematic structure of a body of research.

Figure 11. A Visualized Word cloud of authors' keywords

Figure 12. A Visualized Word cloud of keywords plus field

4.6. Conceptual Structure Map using Multiple Correspondence Analysis

Bibliometrix employs network analysis, correspondence analysis (CA), and multiple correspondence analysis (MCA) to examine phrases found in articles' titles, abstracts, and keywords. CA and MCA visually represent the conceptual organization in separate planes [40]. By utilizing multiple correspondence analysis, a framework of the field is created to identify groups of articles sharing similar concepts, and these findings are displayed on a two-dimensional network. MCA's conceptual structure map encompasses all keywords while considering network uniformity [40].

Figure 13 illustrates the categorization of typical keywords into two distinct groups. The blue cluster encompasses terms related to machine learning and cardiovascular disease, including techniques such as logistic regression, support vector machines, and random forests, as well as applications in cardiology like heart disease diagnosis and classification of information. This cluster highlights a focus on the development and application of computational technologies to improve disease diagnosis, prediction, and classification. In contrast, the red cluster includes keywords associated with the evaluative and methodological aspects of clinical research, such as risk assessment, hypertension, diabetes mellitus, controlled studies, and evaluation metrics like sensitivity and specificity. This cluster emphasizes the concentration

on assessing risk factors and validating diagnostic methods in clinical studies. The interconnections within each cluster, represented by lines of varying thickness, denote the strength and frequency of keyword co-occurrence, with thicker lines indicating stronger associations. The proximity of keywords within the clusters further illustrates their conceptual relatedness, emphasizing the integration of machine learning techniques with cardiology applications in the blue cluster and the methodological rigor of clinical research in the red cluster. This map effectively distills the field into two primary domains: technological advancements and applications of machine learning in cardiology, and clinical evaluation and risk assessment methodologies, highlighting the dual nature of research in this area.

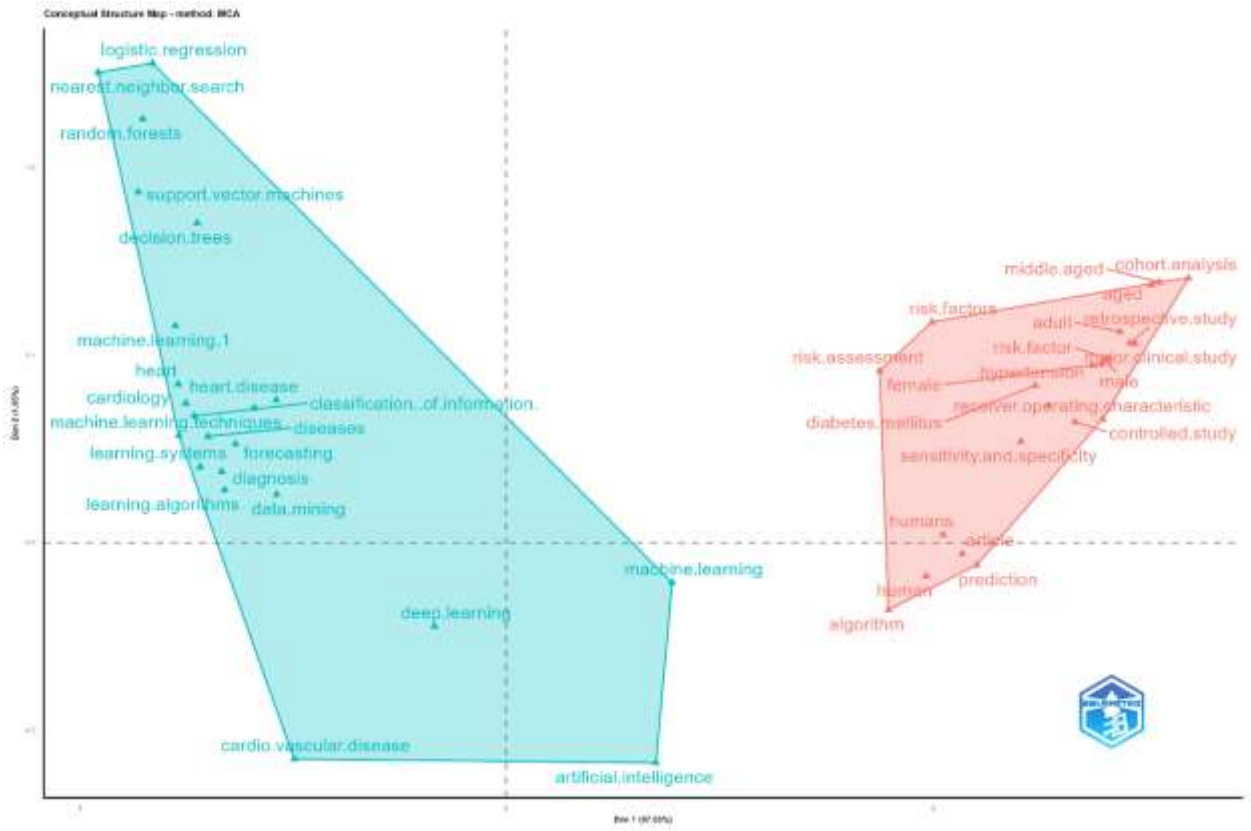

Figure 13. Structure map developed from the multiple correspondence analysis using Biblioshiny software.

4.7. Co-authorship network visualization of countries

The data was utilized to generate a detailed representation of co-authorship density using VOSviewer software. This visualization showcases countries as circles, and the size of each circle corresponds to the number of scientific publications they have contributed to the field of cardiovascular disease prediction using machine learning. The proximity between circles reflects the level of co-authorship collaboration among countries, with closer circles indicating stronger connections.

The visualization was intended to highlight the most prominent countries in terms of publication volume and collaboration in the field. As such, not all 61 countries are displayed with equal prominence; this results in some countries with fewer contributions appearing less visibly or not at all in the co-authorship density map. Figure 14 presents a visual depiction of the data gathered from a study encompassing multiple countries. The data was obtained by tallying the number of publications produced by 126 countries, resulting in a total of 2,564 publications. The minimum and maximum number of countries that contributed to each paper was 5, and 61 countries met the prerequisite.

Through the mapping process, nine clusters were identified, containing a total of 61 items distributed as follows: cluster 1 (12 items), cluster 2 (9 items), cluster 3 (8 items), clusters 4 and 5 (7 items), cluster 6 (6 items), clusters 7 and 8 (5 items), and cluster 9 (2 items). After investigation, it was discovered that the United States had the highest overall link strength, reaching 378, with 427 articles and 8,434 citations. With an absolute link strength of 204 and 120 papers, the United Kingdom came in second. India had the most documents (1,008) with a link strength of 177 and 7,307 citations. The United States appears to be the largest node, implying it has the highest number of collaborations. Other prominent nodes include China, India, and several European countries such as Germany, France, and the United Kingdom. The proximity of nodes to each other might reflect the frequency or intensity of collaborations between these countries.

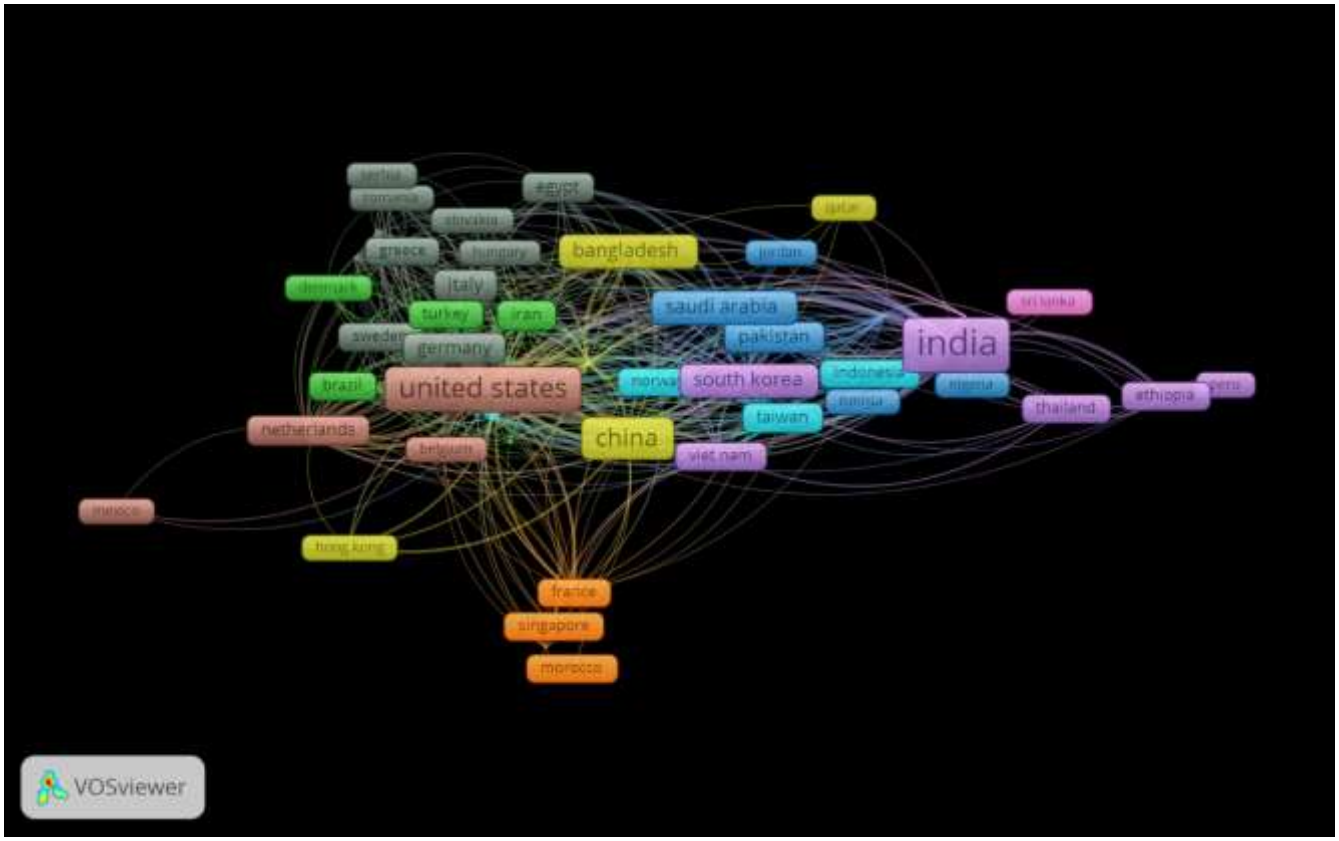

Figure 14. Network visualization of the co-authorship of countries generated using VOSviewer.

4.8. Bibliographic Coupling with Sources

The illustration provides a graphic representation of interconnected sources relevant to the prognosis of cardiovascular disease using machine learning. Of the 1150 sources that published articles on this topic, only 92 met the necessary criteria. These criteria included selecting sources that had published a minimum of five articles and employing a comprehensive full counting method to assess their relevance. By mapping this network, we can visualize the connections and relationships between research articles and their sources. The total strength of the bibliographic coupling links among the 92 sources has been calculated. The highest value obtained for the total link strength (TLS) from these sources was 7090 TLS. Using this value, the sources were divided into six clusters, consisting of 92 items in total. The distribution of items across the clusters is as follows: the first cluster contains 24 items, the second cluster has 19 items, the third and fourth clusters each contain 16 items, the fifth cluster contains 15 items, and the last cluster has two items. Furthermore, the data presented reveals that the highest combined link strength achieved was 1095. This involved 33 articles that re-

ceived a total of 1638 citations from the journal "IEEE Access," positioning it at the top. The journal "Lecture Notes in networks and Systems" followed in second place with 682 combined link strengths from 56 research articles. These findings suggest a significant collaboration between these two journals in publishing academic papers, as depicted in Figure 15.

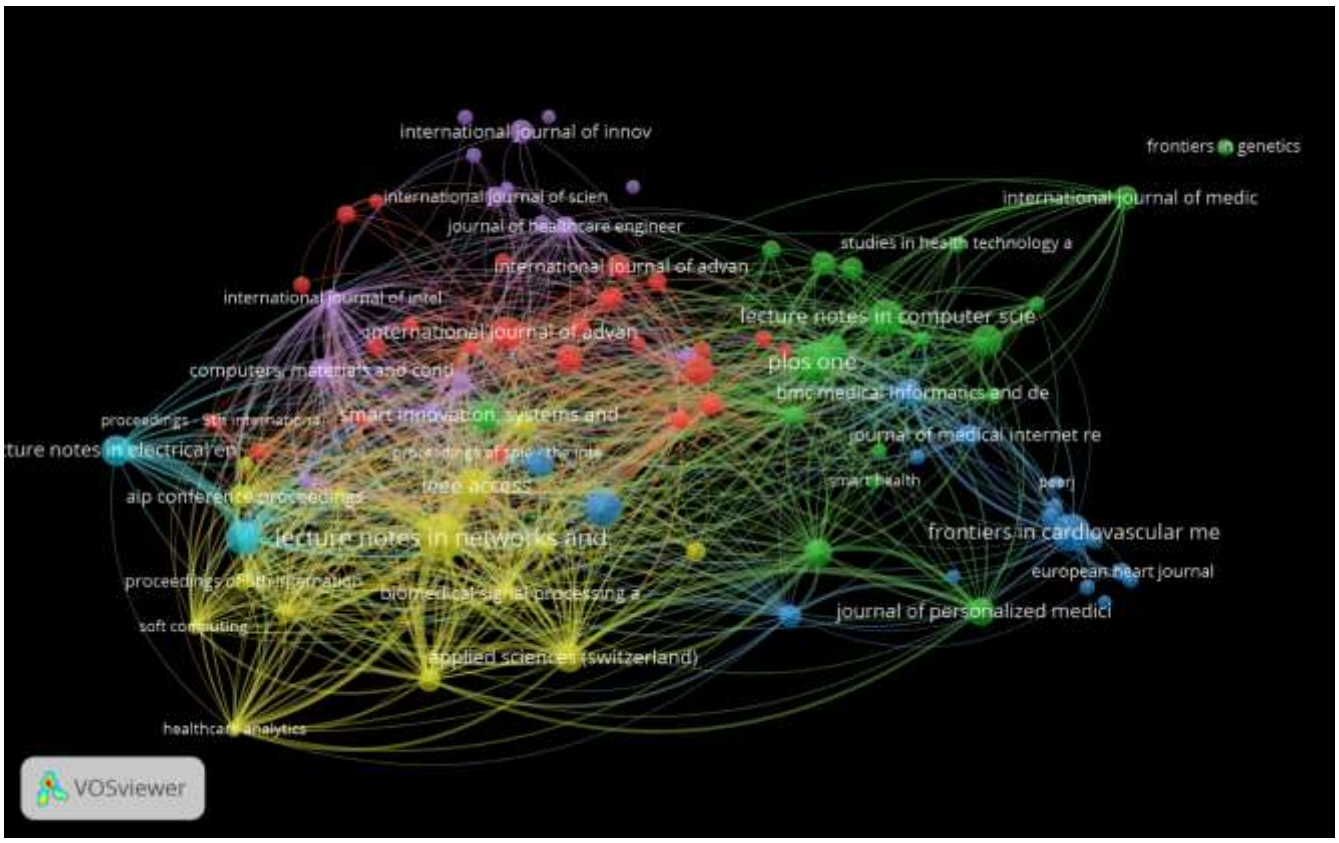

Figure 15. VOSviewer network visualization of the bibliographic coupling with sources.

4.9. Bibliographic Coupling with Countries

Figure 16 illustrates the relationship between bibliographic coupling and various countries in the context of cardiovascular disease prediction using machine learning research. Bibliographic coupling occurs when two research publications reference the same set of prior works, thereby establishing a conceptual link between them. This linkage suggests that the studies are potentially related in their subject matter or methodology. In our visualization, these links are represented by lines connecting country nodes, with the thickness of a line corresponding to the number of shared references, thereby indicating the strength of the bibliographic relationship. A denser cluster of lines between countries signifies a higher degree of intellectual exchange and thematic overlap in the realm of cardiovascular disease prediction using machine learning. This measure provides a proxy for the extent to which researchers in different countries are engaged in dialogues or debates central to this specialized field. A total of 126 countries participated in publishing academic papers, and among them, 62 countries surpassed the threshold of five publications. Consequently, the figure displays nine distinct clusters comprising a combined total of 62 items. Specifically, cluster 1 consists of 18 items, cluster 2 contains 13 items, cluster 3 has 8 items, cluster 4 contains 7 items, clusters 5 and 6 both have 6 items, cluster 7 has 2 items, while clusters 8 and 9 each consist of only one item. Within these clusters, India emerges as the country with the highest number of bibliographic coupling links, amounting to 35,773 occurrences. India achieved this through 1,008 documents that garnered 7,307 citations. The United States follows closely with 26,639 bibliographic coupling links across 427 documents, which received 8,434 citations. With India having the highest number of bibliographic coupling links, it suggests that Indian researchers

are extensively citing a broad range of research that aligns with the work of many other countries. Thus, India would be central to the network, indicating strong collaborative ties with various countries within its cluster. The United States has fewer documents than India, the high number of citations can suggest that its research is highly referenced, potentially indicating quality or impactful research that is recognized and used by other countries. This pattern of bibliographic coupling suggests that there is a substantial degree of research interdependency between India and the United States, as each country frequently builds upon the research findings of the other within the field of cardiovascular disease prediction using machine learning.

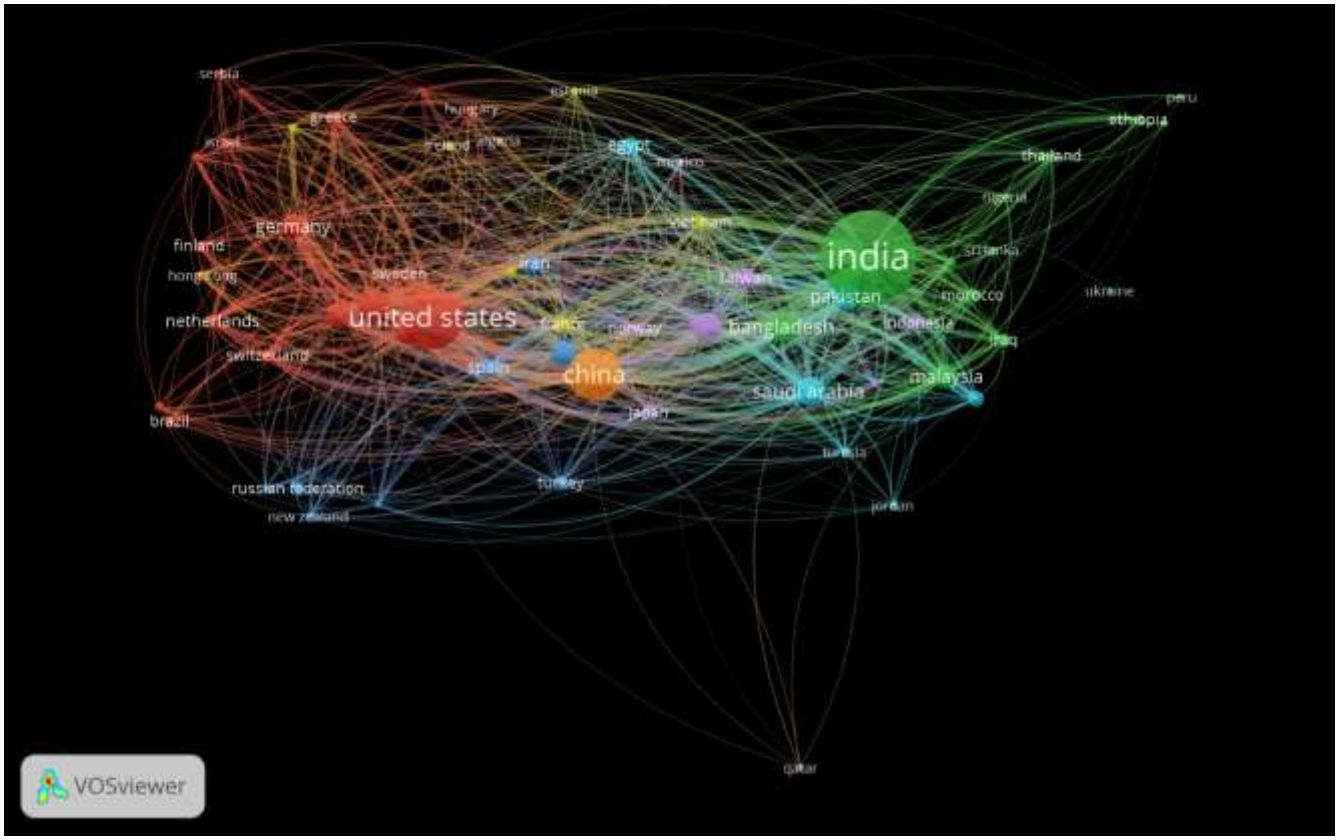

Figure 16. VOSviewer network visualization of the bibliographic coupling with countries.

5. Discussion

A total of 2564 entries were found in the Scopus database search, encompassing publications released between 1991 and 2023. The publication rate exhibited a noteworthy increase starting from 2015 and continued to grow until 2022. This growth in publication count over the years indicates a growing interest and activity in the field covered by the database.

The field of predicting cardiovascular disease using machine learning has a robust research community with numerous prolific authors. The high publication counts and consistent presence of certain authors indicate their influential positions within the field, while the visual representations provide insights into their productivity and impact over time. Two authors, Liu Y and Wang Y, stand out with the highest number of publications, each credited with 25 articles. Liu Y closely follows with 24 articles.

Among the analyzed journals, Lecture notes in networks and systems and Plos One emerged as the most productive, with 56 and 37 publications respectively, indicating their active involvement in publishing research in this field. In terms of institutions involved in this research, Harvard Medical School takes the lead with 60 research publications, closely followed by the University of Oxford with 53 publications. This suggests that these institutions have been actively conducting research on predicting cardiovascular disease. Affiliated institutions Stanford University, Chungbuk National University, and Capital Medical Uni-

versity have also made substantial contributions to the research in this area. The material shows a strong academic interest in cardiovascular disease prediction research, with numerous journals and institutes actively working to advance our understanding and knowledge in this field.

"Machine learning" was found to be the most commonly used keyword in the review of scholarly literature pertaining to the prediction of cardiovascular disease using machine learning. This indicates that machine learning techniques are widespread in studies aimed at predicting cardiovascular disease. The analysis of the most popular authors, phrases, and sources also revealed several well-known concepts, including "random forest," "cardiovascular disease," "heart disease," and "prediction." These concepts are pertinent to the subject matter and extensively covered in the literature. Many well-known authors, including Chen Y, Li Y, Zhang Y, Wang Y, Sharma A, and Gupta A, were also discovered to have published their work in respected journals, including Frontiers in Cardiovascular Medicine and Computers in Biology and Medicine. This indicates that these authors have made significant contributions to the field and their research is well-regarded in the scientific community.

The use of word clouds allows researchers to visualize and prioritize the most important terms, enabling them to focus on specific areas of interest, such as machine learning in cardiology and forecasting of diseases, including cardiovascular and heart diseases. The categorization of keywords using Multiple Correspondence Analysis into two clusters may indicate that there are different themes or areas of focus within the broader topic being discussed. Cluster 1 seems to revolve around machine learning techniques applied to medical fields, particularly in relation to cardiovascular diseases and diagnosis. Cluster 2 encompasses a broader range of methodologies, including both statistical and computational approaches, rather than solely focusing on traditional statistical analysis and study design.

The United States, United Kingdom, and India were prominent contributors in terms of link strength, document count, and citations. Additionally, the United States stood out as a leading collaborator in research on cardiovascular disease prediction. There is a network of interconnected sources focused on cardiovascular disease prediction. The clustering and distribution of items indicate varying degrees of collaboration and importance among the sources. The dominance of "IEEE Access" and the collaboration with "Lecture Notes in Networks and Systems" suggest the influence of these journals in publishing academic papers related to cardiovascular disease prediction using machine learning.

India and the United States have a strong collaboration in cardiovascular disease prediction research. According to the statistics, India has the most bibliographic coupling relationships, demonstrating a strong connection between Indian researchers and those in other nations. Given that they have the most bibliographic coupling relationships out of all the nations discussed, India and the United States' cooperation particularly stands out. Additionally, the data shows that India and the United States have produced an enormous number of scholarly works in this area, with India contributing 1,008 documents and the United States contributing 427 documents. These articles have also attracted an impressive number of citations indicating their significance and impact on the scholarly community.

6. Conclusion

The paper "Visualizing the Impact of Machine Learning on Cardiovascular Disease Prediction: A Comprehensive Analysis of Research Trends" provides information on the advancement and significance of machine learning-based heart disease prediction research. The study offers vital insights into the development of the field, research themes, significant publications and authors, collaboration networks, and research impact through a detailed bibliometric analysis. According to the study, there has been a noticeable rise in papers discussing machine learning-based cardiac disease prediction, showing a growing need for this type of study. The analysis shows the numerous areas researchers have concentrated on by categorizing the research subjects, including feature selection, classification techniques, data prepro-

cessing, and model evaluation. Researchers can use this information to help them decide which crucial topics to investigate further. The study also reveals collaboration networks among sources and nations, highlighting the value of multidisciplinary study and knowledge exchange. For academics, practitioners, and policymakers, its findings provide insightful information that helps identify essential study areas, unique works, collaborative opportunities, and viable directions for further development of machine learning algorithms for heart disease prediction.

References

1. Centers for Disease Control and Prevention (CDC). (2020). Deaths:leading causes. https://www.cdc.gov/nchs/fastats/leading-causes-of-death
2. Kiran, S., Reddy, G. R., S. P., G., Dorthi, K., & V. (2023). A Gradient Boosted Decision Tree with Binary Spotted Hyena Optimizer for cardiovascular disease detection and classification. *Healthcare Analytics*, *3*, 100173. https://doi.org/10.1016/j.health.2023.100173
3. Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, Ruinan. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, *2018*, article ID 3860146. https://doi.org/10.1155/2018/3860146
4. Mohapatra, S., Maneesha, S., Mohanty, S., Patra, P. K., Bhoi, S. K., Sahoo, K. S., & Gandomi, A. H. (2023). A stacking classifiers model for detecting heart irregularities and predicting cardiovascular Disease. *Healthcare Analytics*, *3*, 100133. https://doi.org/10.1016/j.health.2022.100133
5. Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H., & Raad, A. (2023). Reviewing multimodal machine learning and its use in cardiovascular diseases detection. *Electronics*, *12*(7), 1558. https://doi.org/10.3390/electronics12071558
6. Chandrasekhar, N., & Peddakrishna, S. (2023). Enhancing heart disease prediction accuracy through machine learning techniques and optimization. *Processes*, *11*(4), 1210. https://doi.org/10.3390/pr11041210
7. Houssein, E. H., Mohamed, R. E., & Ali, A. A. (2023). Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques. *Scientific Reports*, *13*(1), 7173. https://doi.org/10.1038/s41598-023-34294-6
8. Biswas, N., Ali, M. M., Rahaman, M. A., Islam, M., Mia, M. R., Azam, S., Ahmed, K., Bui, F. M., Al-Zahrani, F. A., & Moni, M. A.. (2023). Machine learning-based model to predict heart disease in early stage employing different feature selection techniques. *BioMed Research International*, *2023*, article ID 6864343. https://doi.org/10.1155/2023/6864343
9. Thompson, D. F., & Walker, C. K. (2015). A descriptive and historical review of bibliometrics with applications to medical sciences. *Pharmacotherapy*, *35*(6), 551–559. https://doi.org/10.1002/phar.1586
10. Zhang, T., Yin, X., Yang, X., Man, J., He, Q., Wu, Q., & Lu, M. (2020). Research trends on the relationship between microbiota and Gastric Cancer: A bibliometric analysis from 2000 to 2019. *Journal of Cancer*, *11*(16), 4823–4831. https://doi.org/10.7150/jca.44126
11. Li, R., Liu, B., Yuan, X., & Chen, Z. (2023). A bibliometric analysis of research on R-loop: Landscapes, highlights and trending topics. *DNA Repair*, *127*, 103502. https://doi.org/10.1016/j.dnarep.2023.103502
12. Muda, M. F., Hashim, M. H. M., Rahman, A., Mohd, M. H., Khairul, M., Al-Fakih, A., Haza, Z. F., & Sam, R. M. (2023). A Trend in Pipeline Rehabilitation Research: A Bibliometric Analysis. 18.
13. Cai, X. J., Zhang, H. Y., Zhang, J. Y., & Li, T. J. (2023). Bibliometric analysis of immunotherapy for head and neck squamous cell carcinoma. *Journal of Dental Sciences*, *18*(2), 872–882. https://doi.org/10.1016/j.jds.2023.02.007
14. Mohd Sofian, F. N. R., Abdullah, K. H., & Mohd-Sabrun, I. (2023). Research on corporate reputation: A bibliometric review of 43 years (1977–2020). International Journal of In-

formation Science and Management (IJISM), 21(2), 31–54. https://doi.org/10.22034/ijism.2023.1977558.0

15. Wan, G., Yahya Dawod, A., & Nopasit, C. (2023). A bibliometric and visual analysis in the field of environment, social and governance (ESG) between 2004 and 2021. International Journal of Information Science and Management (IJISM), 21(2), 103–125. https://doi.org/10.22034/ijism.2023.1977765.0

16. Van Eck, N. J., & Waltman, L. (August 2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

17. Utami, N., Setiawan, A., & Hamidah, I. (2023). A Bibliometric Analysis of Augmented Reality in Higher Education. 18.

18. Yu, Y., Li, Y., Zhang, Z., Gu, Z., Zhong, H., Zha, Q., Yang, L., Zhu, C., & Chen, E. (July 2020). A bibliometric analysis using VOSviewer of publications on COVID-19. *Annals of Translational Medicine*, *8*(13), 816–816. https://doi.org/10.21037/atm-20-4235

19. McAllister, J. T., Lennertz, L., & Atencio Mojica, Z. A. (2022). Mapping A discipline: A guide to using VOSviewer for bibliometric and visual analysis. *Science and Technology Libraries*, *41*(3), 319–348. https://doi.org/10.1080/0194262X.2021.1991547

20. Aria, M., & Cuccurullo, C. (November 2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

21. Maryanti, R., Nandiyanto, A. B. D., Hufad, A., Sunardi, S., Husaeni, D. N. A., & Husaeni, D. F. A. (2023). A Computational Bibliometric Analysis of Science Education Research Using VOSviewer. 18.

22. Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 281. https://doi.org/10.1186/s12911-019-1004-8

23. Patro, S. P., Nayak, G. S., & Padhy, N. (2021). Heart disease prediction by using novel optimization algorithm: A supervised learning prospective. *Informatics in Medicine Unlocked*, *26*, 100696. https://doi.org/10.1016/j.imu.2021.100696

24. Gupta, A., Kumar, R., Singh Arora, H., & Raman, B. (2020). MIFH: A machine intelligence framework for heart disease diagnosis. *IEEE Access*, *8*, 14659–14674. https://doi.org/10.1109/ACCESS.2019.2962755

25. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

26. Shah, S. M. S., Batool, S., Khan, I., Ashraf, M. U., Abbas, S. H., & Hussain, S. A. (2017). Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis. *Physica A*, *482*, 796–807. https://doi.org/10.1016/j.physa.2017.04.113

27. Perumal, R. (2020). Early prediction of coronary heart disease from Cleveland dataset using machine learning techniques. *International Journal of Advanced Science and Technology*, *29*, 4225–4234.

28. UCI Heart Disease Data set [Online]. https://archive.ics.uci.edu/ml/datasets/heart+disease

29. Statlong Heart Data set [Online]. https://archive.ics.uci.edu/ml/datasets/statlog+(heart)

30. Verma, L., Srivastava, S., & Negi, P. C. (2016). A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. *Journal of Medical Systems*, *40*(7), 178. http://doi.org/10.1007/s10916-016-0536-z

31. M. (2020). Siddhartha, heart disease dataset (comprehensive), IEEE Dataport. https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive#files

32. Alalawi, H. H., & Alsuwat, M. S. (2021). Detection of cardiovascular disease using machine learning classification models. *International Journal of Engineering Research and Technology*, *10*(07), 151–157.

33. Yılmaz, R., & Yağın, F. H. (2022). Early detection of coronary heart disease based on machine learning methods. *Medical Records*, *4*(1), 1–6. http://dx.doi. https://doi.org/10.37990/medr.1011924

34. Pires, I. M., Marques, G., Garcia, N. M., & Ponciano, V. (2020). Machine learning for the evaluation of the presence of heart disease, Procedia Comput. Sci, *177*, 432–437. http://doi.org/10.1016/j.procs. 2020.10.058. (ISSN. 1877, *0509*).

35. Subramani, S., Varshney, N., Anand, M. V., Soudagar, M. E. M., Al-Keridis, L. A., Upadhyay, T. K., Alshammari, N., Saeed, M., Subramanian, K., Anbarasu, K., & Rohini, K. (2023). Cardiovascular diseases prediction by machine learning incorporation with deep learning. *Frontiers in Medicine*, *10*, 1150933. https://doi.org/10.3389/fmed.2023.1150933

36. Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, *128*, 102289. https://doi.org/10.1016/j.artmed.2022.102289

37. Brites, I. S., da Silva, L. M., Barbosa, J. L., Rigo, S. J., Correia, S. D., & Leithardt, V. R. (2021). Machine Learning and IoT Applied to Cardiovascular Diseases Identification through Heart Sounds: A Literature Review. *Informatics*, *8*(4). https://doi.org/10.3390/informatics8040073

38. Riehmann, P., Hanfler, M., & Froehlich, B. (October 23–25, 2005). Interactive Sankey diagrams. In *Proceedings of the IEEE Symposium on Information Visualization* (pp. 233–240). INFOVIS. https://doi.org/10.1109/INFVIS.2005.1532152

39. Kumar, R., Singh, S., Sidhu, A. S., & Pruncu, C. I. (2021). Bibliometric analysis of specific energy consumption (SEC) in machining operations: A sustainable response. *Sustainability*, *13*(10), 5617. https://doi.org/10.3390/su13105617

40. B M, L., Chakraborty, S., Kumar Ghosh, B., & Shenoy U, R.. (2021). Overview of bond mutual funds: A systematic and bibliometric review. *Cogent Business and Management*, *8*(1), 1, 1979386. https://doi.org/10.1080/23311975.2021.1979386