OKU Fen Bilimleri Enstitüsü Dergisi
8(2): 632-651, 2025

OKU Journal of The Institute of Science and Technology, 8(2): 632-651, 2025

Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi

Osmaniye Korkut Ata University Journal of The Institute of Science and Technology

Varyant Etki Tahmininde Makine Öğrenmesi Yöntemlerinin Kullanımına Yönelik Bibliyometrik Bir Analiz

Gülbahar Merve ŞILBIR[1*], Burçin KURT[2]

[1]Trabzon Üniversitesi, Trabzon
[2]Karadeniz Teknik Üniversitesi, Tıp Fakültesi, Biyoistatistik ve Tıp Bilişimi Bölümü, Trabzon

[1]https://orcid.org/0000-0003-0321-7259
[2]https://orcid.org/0000-0001-5781-2382
*Sorumlu yazar: gmervecakmak@trabzon.edu.tr

Araştırma Makalesi

Makale Tarihçesi:
Geliş tarihi: 27.06.2024
Kabul tarihi: 12.12.2024
Online Yayınlanma: 12.03.2025

Anahtar Kelimeler:
Bibliyometrik analiz
Bibliometriks
Varyant etki tahmini
Makine öğrenmesi

ÖZ

İnsan genomunda oluşan varyantların fenotip üzerindeki etkisinin tahmin edilmesinde yapay zeka, makine öğrenmesi ve derin öğrenme gibi hesaplamalı yöntemlerin kullanıldığı çalışmalar son zamanlarda giderek artmaktadır. Bu çalışmanın amacı bibliyometrik yöntem kullanılarak varyant etki tahmininde makine öğrenmesi yöntemlerinin kullanıldığı bilimsel araştırmalara genel bir bakış sunmaktır. Bu amaç doğrultusunda çalışmada ilgili literatüre ulaşmak için Web of Science Core Collection (WoSCC) veritabanı kullanılmıştır. Ülkeler, kurumlar, yazarlar, dergiler, alıntılar ve anahtar kelimeler R-Studio programında "bibliometrix" kütüphanesi kullanılarak analiz edilmiştir. Yapılan analiz sonucunda göre varyant etki tahmininde makine öğrenmesi yöntemlerinin kullanımına ilişkin yapılan bilimsel yayınların son yıllarda popülerliğinin giderek arttığı ve bu artıştaki en büyük payın United States'te yer alan kurumların China, Germany, England ve Australia ile ortak araştırmalara bağlı olduğu görülmüştür. En çok atıf alan yazarın Jian Zhou (1.116), yazarlar arası atıflarda öne çıkan yazarların ise Jonathan Frazer ve José Juan Almagro Armenteros olduğu görülmüştür. Çalışmalarda ele alınan konuların makine öğrenmesi ve derin öğrenme temalarında şekillendiği belirlenmiştir. Bu alanda araştırılan konular arasında amino asit varyantları, genom düzeyindeki mutasyonlar, varyantların yapısal bilgileri, covid-19 mutasyonları ve protein yapısı yer almaktadır. Gelecekteki çalışmalarda bu araştırma konuları makine öğrenmesi ve derin öğrenmeye dayalı farklı yöntemlerle araştırılabilir.

A Bibliometric Analysis of the Use of Machine Learning Methods in Variant Effect Prediction

Research Article

Article History:
Received: 27.06.2024
Accepted: 12.12.2024
Published online: 12.03.2025

Keywords:
Bibliometric analysis
Bibliometrix
Variant effect prediction
Machine learning

ABSTRACT

The application of computational methods, including artificial intelligence, machine learning, and deep learning, to the prediction of the effects of variants in the human genome on phenotype has been on the rise in recent times. The objective of this study is to present a comprehensive overview of scientific studies utilizing machine learning methodologies for the prediction of variant effects, employing the bibliometric approach. To achieve this, the Web of Science Core Collection (WoSCC) database was utilized to access the relevant literature pertinent to the study. The analysis was conducted using the "bibliometrix" library in the R-Studio program, with a focus on countries, institutions, authors, journals, citations, and keywords. The results of the analysis indicate that the popularity of scientific publications on the use of machine learning methods in variant effect prediction has increased in recent years. This growth can be attributed primarily to collaborative research efforts

between institutions in the United States and those in China, Germany, England, and Australia. The most frequently cited author was Jian Zhou, with 1.116 citations. Jonathan Frazer and José Juan Almagro Armenteros were the most prominent authors in terms of citations between authors. The studies revealed that the topics covered were shaped by the themes of machine learning and deep learning. The topics researched in this field included amino acid variants, mutations at the genome level, structural information of variants, covid-19 mutations, and protein structure. In future studies, these research topics can be investigated with different methods based on machine learning and deep learning.

To Cite: Şılbır GM., Kurt B. A Bibliometric Analysis of the Use of Machine Learning Methods in Variant Effect Prediction. Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi 2025; 8(2): 632-651.

1. Introduction

A significant advancement has been made in the field of bioinformatics with the implementation of large-scale genome projects. These projects examine the human genome structure to identify genes associated with diseases (The International HapMap Consortium, 2003; The ENCODE Project Consortium, 2007; The 1000 Genomes Project Consortium, 2010; Fidanoğlu et al., 2013; Qu and Fang, 2013). Some changes in the genome, which play a fundamental role in the formation of differences between individuals, may occur at different frequencies and structures (Tang and Thomas, 2016). Should these changes occur in a manner that may give rise to disease in the human phenotype, it becomes imperative to conduct experimental or computational studies at the genome level to gain further insight into the pertinent alteration (Niroula and Vihinen, 2016; Xu et al., 2021). Some studies have been conducted to determine the pathogenicity effects of these changes, which are known as genetic variants, on the proteins that are the products of the genes in question, due to their potential to directly affect the phenotype (Qiu et al., 2020). In particular, studies employing computational methods, including artificial intelligence, machine learning, and deep learning, have gained prominence in recent years as a means of determining the impact of these millions of variants on the phenotype, as evidenced by whole genome sequencing studies (Li et al., 2012; Ionita-Laza et al., 2016; Livesey and Marsh, 2023).

In the process of predicting the effect that a genetic variant may have on the phenotype, three principal types of information are employed: Information about protein sequence, evolutionary conservation, and structural characteristics is utilized in this process (Tang and Thomas, 2016; Riesselman et al., 2018). The number of studies in which this information is obtained from open-access databases and analyzed with computational methods to predict whether the variant is pathogenic is increasing daily. However, obtaining this data for each of the millions of variants can be a laborious and time-consuming process (Angermueller et al., 2016). Consequently, it is more feasible to develop predictions about variants with unknown effects based on variants with experimentally proven effects. In this regard, the potential of machine learning methods in variant effect prediction is being investigated.

A substantial corpus of scientific literature exists on the subject of machine learning and its application to the prediction of variant effects. The studies in this literature employ a variety of variant datasets and computational methods (Rentzsch et al., 2021; Horne and Shukla, 2022; Bromberg et al., 2024). Mahmood et al. (2017) compared the performance of the developed prediction models on different

benchmark datasets in their study in which they evaluated the success of the developed prediction models in their publications investigating the use of machine learning methods in variant effect prediction. Similarly, Niroula and Vihinen (2019) tried to determine the most successful prediction model for variant effect estimation in the literature. Such evaluation studies appear to provide valuable guidance in determining the most successful model for variant effect prediction. Nevertheless, the extant literature on the utilization of machine learning techniques in variant effect prediction remains incomplete, and the number of generalizations derived from research is limited. To address these deficiencies, the objective of this study is to present the intellectual structure and development over time of scientific publications on the use of machine learning methods in variant effect prediction in the literature. This study aims to provide a source of information with general trends and recommendations to those investigating the use of machine learning methods in variant effect prediction.

This study presents a bibliometric analysis of scientific research utilising machine learning methods for variant effect prediction. The analysis is conducted from a broad perspective, with a particular focus on variant effect prediction from a holistic standpoint. We believe that this study will make a substantial contribution to the extant literature on this subject, providing insights that can inform future scientific studies on the use of machine learning in variable effect prediction.

2. Material and Methods

2.1. Data Source and Search Method

The literature data on the utilisation of machine learning methodologies for the prediction of variant effects was obtained through the utilisation of the Web of Science Core Collection (WoSCC) database. The relevant literature data was published in the Web of Science Core Collection database on 7.3.2024 with the keywords “((variant prediction model or mutation prediction model) same variant effect prediction*) and (machine learning* or deep learning* or supervised learning*) and “topic (topic)”. 335 publications were reached in the research conducted by selecting ")". According to years, 216 articles, 78 Early Access, 19 proceeding papers, 10 review articles and 12 other publication types were accessed, with the oldest being 1995 and the newest being 2024. Filtering procedures were carried out for these publications obtained within the scope of the research. Accordingly, 12 publications (book chapter, data paper, retracted publication) were excluded from the research. The analytical procedures employed in this study are based on the measures proposed by Donthu et al. (2021). The work of Donthu et al. (2021) represents a significant contribution to the field of bibliometric analysis, offering a comprehensive overview of current techniques and procedures. The methodology employed for the identification and analysis of pertinent literature is illustrated in Figure 1.

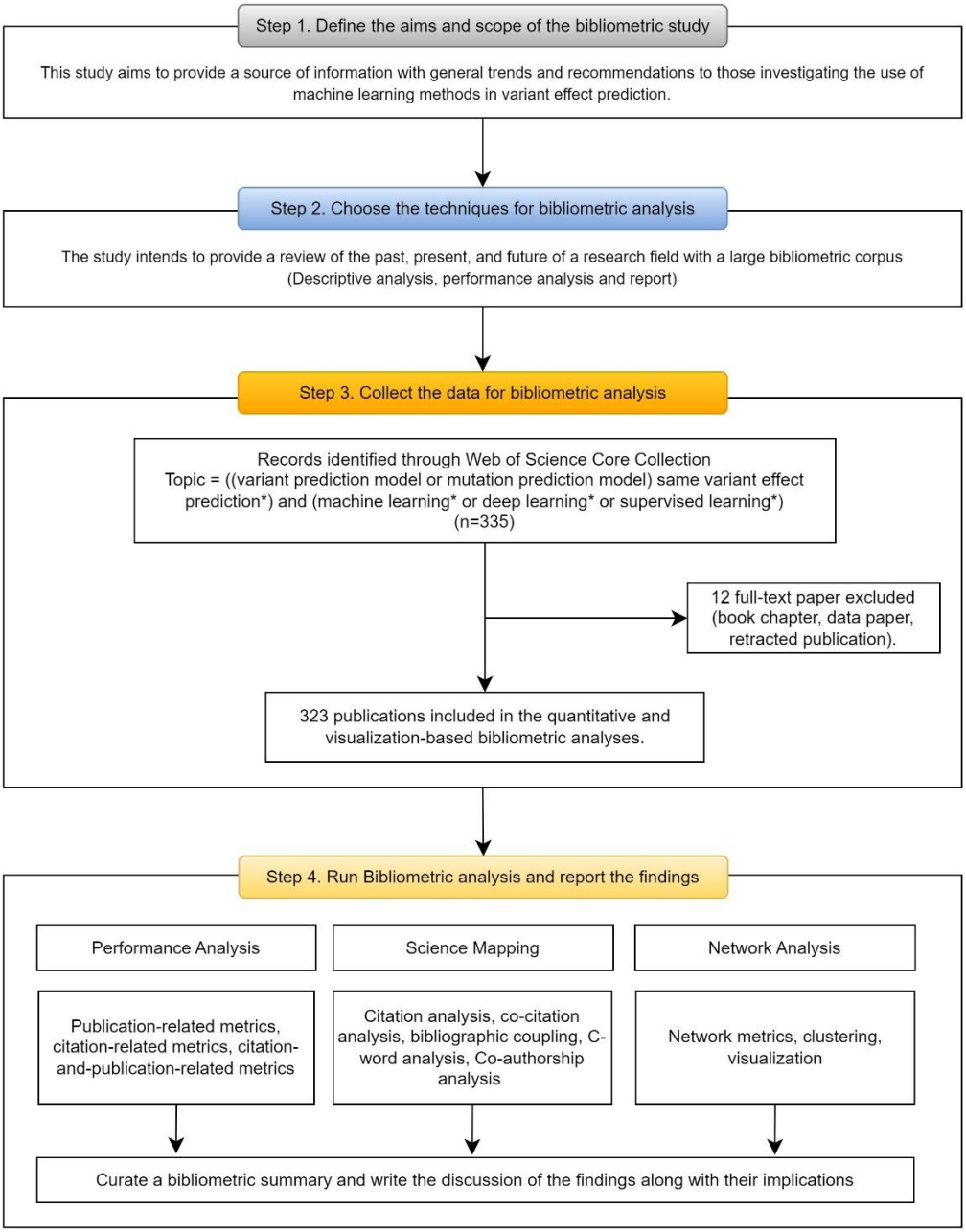

Figure 1. Flow diagram of the publication search and analysis process

2.2. Analysis of Data

To facilitate the research process, the data sources obtained from the WoSCC database were exported as a BibTeX file. The "bibliometrix" library was employed in the R-Studio program to analyse the data (Aria and Cuccurullo, 2017). The distribution of studies on the use of machine learning methods in variant impact estimation research by years, countries and authors, the average number of citations, journals that publish the most on the subject, authors who do the most research, citation percentages of authors, scientific productivity of countries, most cited research documents, collaboration networks of researchers, word cloud maps, co-word analysis, trend topic (keywords and abstact), cumulative distribution of keywords, factor analysis, thematic change were accessed.

3. Results

3.1. Article Distribution by Publication Year

A review of the literature reveals that the use of machine learning methods for variant effect prediction first emerged in 1995. Figure 2 illustrates that a total of 323 articles were published between 1995 and March 2024, with an increasing trend evident in this field on an annual basis. Notably, no scientific articles were published in this field between 1996 and 2003. However, publications investigating the use of machine learning methods in research in the field of variant effect prediction have increased in the following years, particularly after 2017.

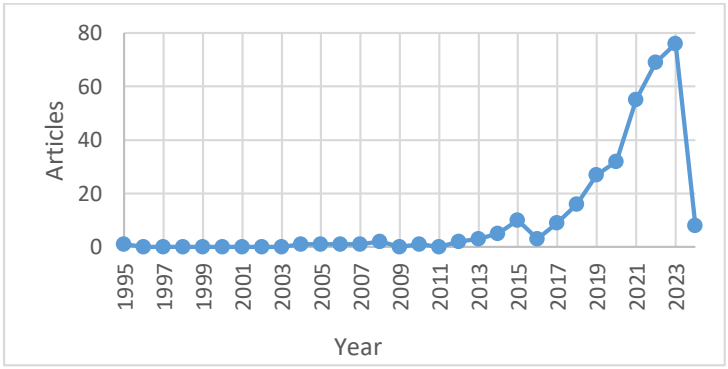

Figure 2. Trends in the number of publications from 1995 to 2024

3.2. Institutes, Countries, and Regions

A review of publications by country reveals that the United States is the leading contributor, with 132 articles, followed by China (69), Germany (34), England (31), India (21), Australia (19), Italy (18), France (14), Canada (13), and Denmark (13). These countries are listed in descending order of publication output. The map in Figure 3 illustrates the international collaboration in research and publication activities among these countries. Upon examining the map in Figure 3, it is observed that the United States is represented in dark blue. This highlights numerous publications showing that the United States plays a critical role in significantly contributing to studies using machine learning methods for variant impact prediction. The frequency of other research in this area is illustrated by a transition from dark blue to light blue on the map displayed in Figure 3. Furthermore, cross-country collaborations are evident in Figure 3.

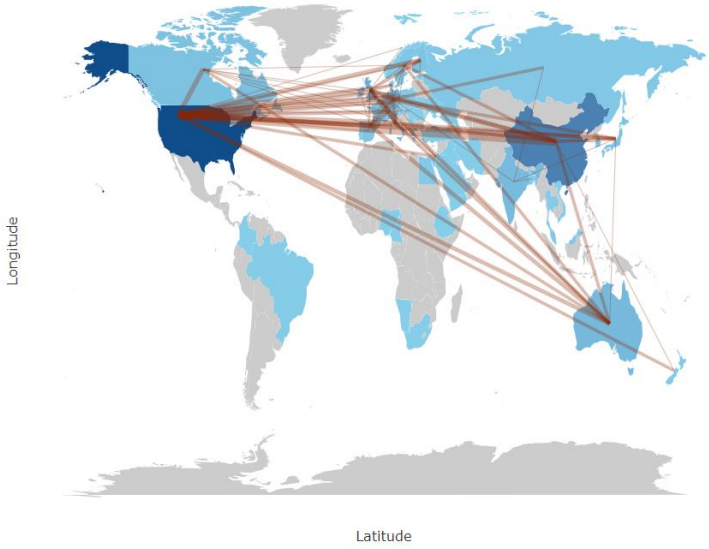

Figure 3. Country collaboration map

Figure 4 presents a ranking of the top ten academic institutions that have made significant contributions to the field of machine learning-based variant impact prediction. The University of California San Francisco, the University of Washington, and Columbia University have emerged as the leading institutions, with respective publication counts of 34, 29, and 22.

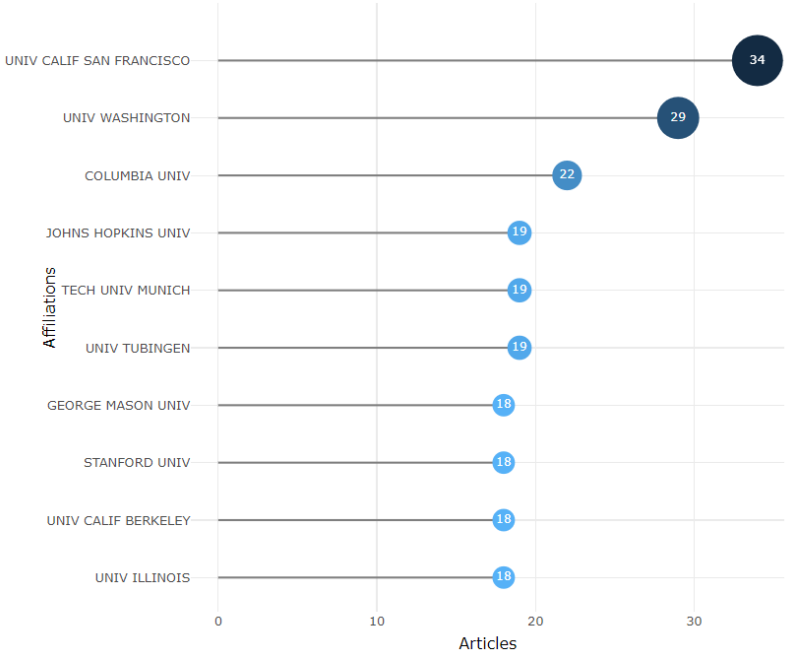

Figure 4. Top 10 relevant institutes

Figure 5 illustrates the countries most frequently cited in scientific publications on the utilisation of machine learning methodologies for variant impact prediction. It can be observed that the United States is the most frequently cited country (2703), followed by Denmark (761), China (654), Germany (555)

and Spain (434). The remaining countries are listed in Figure 5 in descending order of citation, totalling 15 countries.

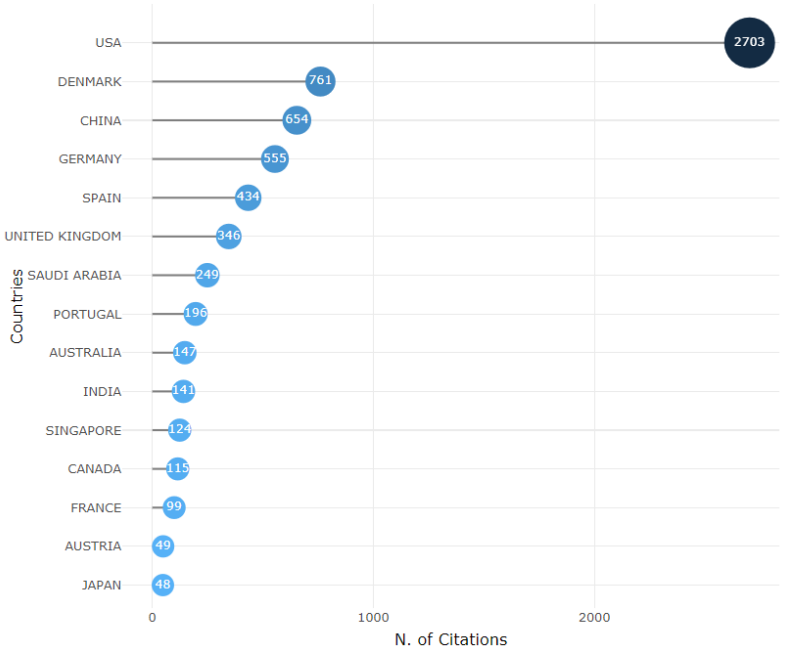

Figure 5. Most cited countries

3.3. Journals

The journals in which scientific research on the use of machine learning methods in variant effect estimation was published were subjected to analysis, and the journals in which the articles in this field were published most frequently were identified. The ten journals that have published the most articles in this field are listed in Figure 6. The most frequently cited journals are Human Mutation (13), BMC Bioinformatics (11), Genome Biology (8), PLOS One (8), and PLOS Computational Biology (7).

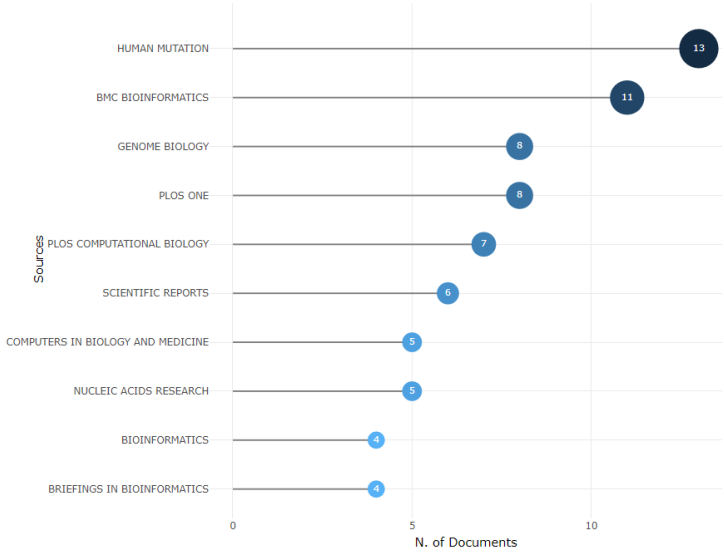

Figure 6. Most relevant sources

By examining the citation levels of the journals in which the relevant publications were published, it is possible to identify the ten journals with the highest number of citations. As illustrated in Figure 7, the most frequently cited journal in this field is Nucleic Acids Research, with a total of 758 citations. Other highly-cited journals include Bioinformatics (581 citations), Nature (463 citations), and Nature Genetics (387 citations).

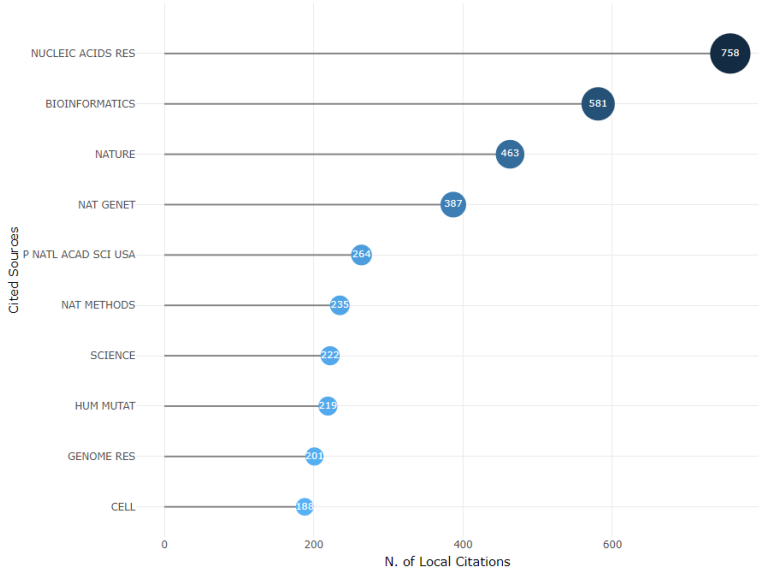

Figure 7. Most local cited sources

3.4. Authors

This study examines the publication frequencies of authors engaged in scientific research on the use of machine learning methods in variant effect prediction. The ten authors with the highest publication counts are presented in Figure 8. The data reveal that Yongguo Liu, Yun Zhang, Majid Masso, Haicang Zhang, Yuedong Yang, and Jiajing Zhu are the most prolific authors in this field, having published 11, 8, 7, 6, and 5 articles, respectively.

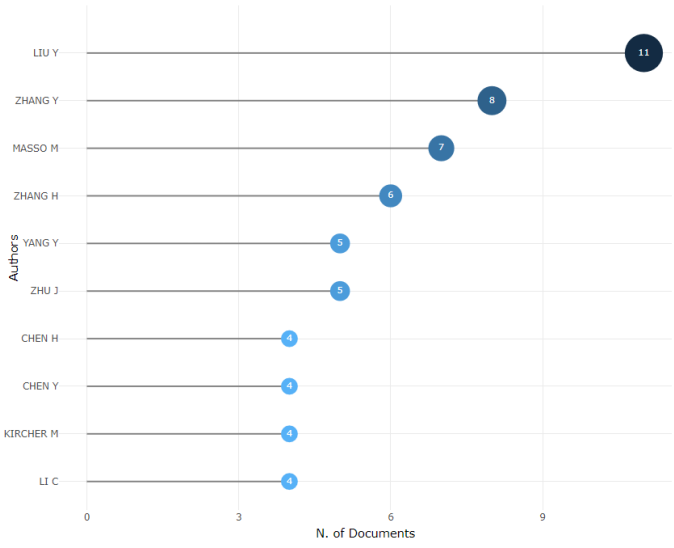

Figure 8. Most relevant authors

The graph in Figure 9 illustrates the evolution of publication output by authors over time. It was observed that the authors in question produced publications between the specified years. The period of publication for Yongguo Liu is 2017-2023, for Yun Zhang it is 2021-2023, and for Majid Masso it is 2008-2020. The graph also shows that Haicang Zhang published between 2019 and 2024, Yuedong Yang between 2013 and 2022, and Jiajing Zhu between 2020 and 2023. It can be seen that Majid Masso and Yuedong Yang have been publishing in this field for a considerable period.

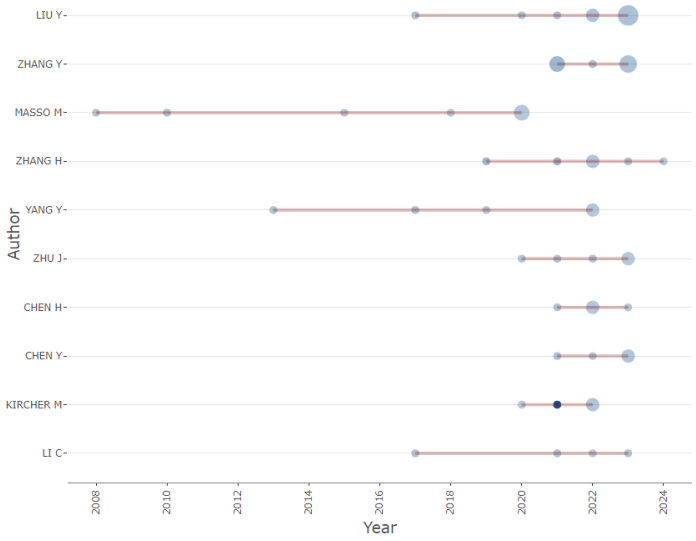

Figure 9. Authors' production over time

The citation levels of authors were examined according to publication year and published journal. The most cited author, publication year and published journal are presented in Figure 10. Upon examination of Figure 10, it can be observed that the article published by Jian Zhou (1.116) in the Nature Methods journal in 2015 received the highest number of citations. The other top ten most cited authors in the field are shown in Figure 10.

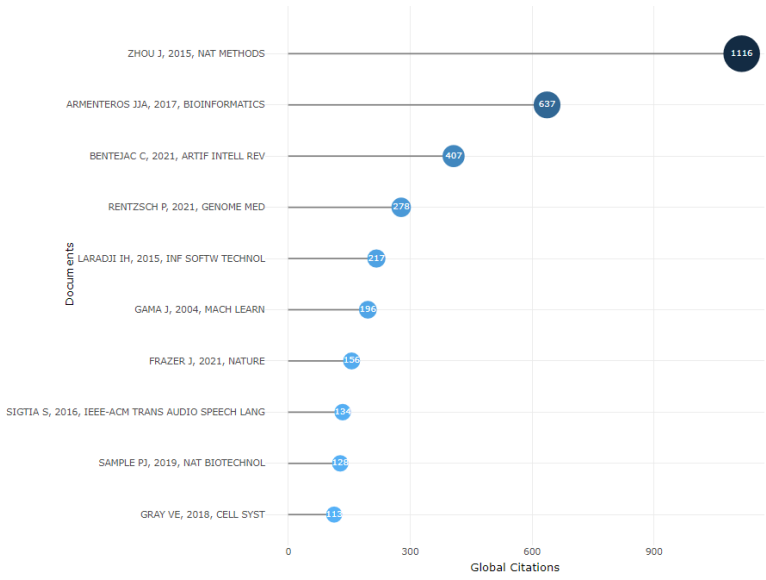

Figure 10. Most globally cited documents

The citation network and the relationships between citations are illustrated in Figure 11. Upon examination of Figure 11, it becomes evident that the most cited authors are represented by coloured circles, with the relationships between them indicated by coloured lines. In this context, Martin Kircher's publication in 2014, Ivan A. Adzhubei's publication in 2010, and Jian Zhou's publication in 2015 emerge as particularly noteworthy.

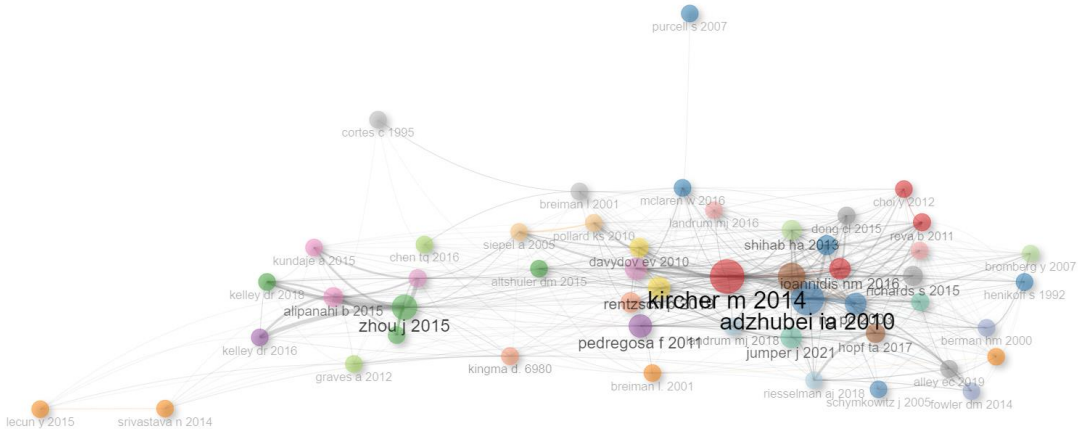

Figure 11. Co-citation network

A Local Citation Score (LCS) and Global Citation Score (GLC) analysis of between-authors citations was conducted, and the resulting network is presented in Figure 12. Upon examination of the network depicted in Figure 12, it becomes evident that the citations between authors are grouped into six clusters, each distinguished by a distinct colour. The lines representing the citations within each cluster are indicated by the same colour as the cluster itself. Table 1 presents the publications belonging to each cluster, along with their respective LCS and GLC values. The LCS value indicates the number of citations made to publications within a given cluster by other publications within that same cluster. In contrast, the GLC value represents the total number of citations a given publication has received.

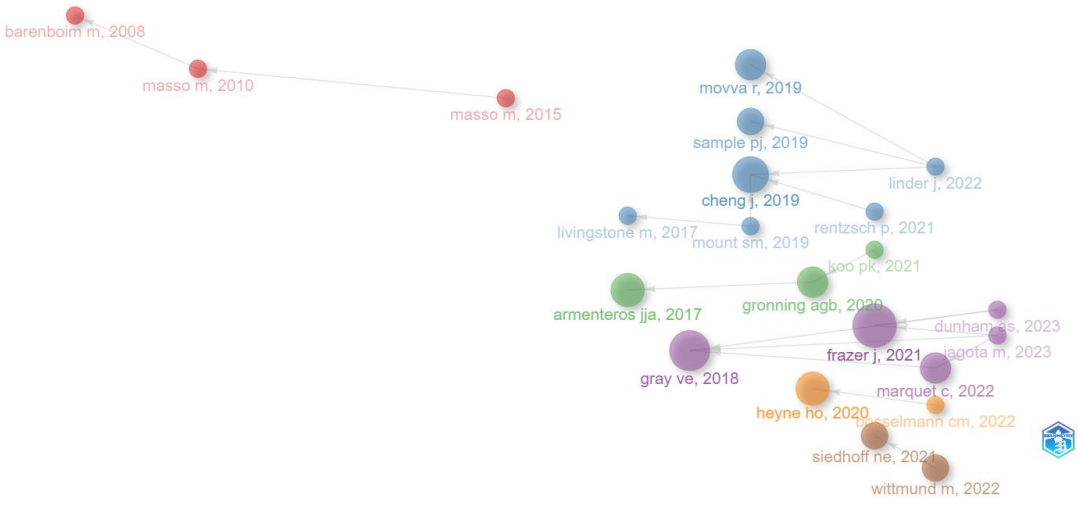

Figure 12. Historiograph

Table 1. Details of papers citation network (historiograph)

Paper	Title	LCS	GCS	Cluster	Color
Barenboim M, 2008, Proteins	Statistical geometry based prediction of nonsynonymous snp functional effects using random forest and neuro-fuzzy classifiers	1	27	1	Red
Masso M, 2010, j theor biol	Knowledge-based computational mutagenesis for predicting the dissease potential of human non-synonymous single nucleotide polymorphisms	1	46	1	
Masso M, 2015, peerj	Modeling functional changes to Escherichia coli thymidylate synthase upon single residue replacements: a structure-based approach	1	3	1	
Livingstone M, 2017, Hum Mutat	Investigating DNA, RNA, and protein-based features as means to discriminate pathogenic synonymous variants	1	30	2	Blue
Mount SM, 2019, Hum Mutat	Assessing predictions of the impact of variants on splicing in CAGI5	1	10	2	
Sample PJ, 2019, Nat Biotechnol	Human 5′ utr design and variable effect prediction from a massively parallel translation assay	2	128	2	
Cheng J, 2019, Genome Biol	Mmsplice: modular modeling improves the predictions of genetic variable effects on splicing	5	95	2	
Movva R, 2019, Plos One	Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays	3	40	2	
Rentzsch P, 2021, Genome Med	Cadd-splice-improving genome-wide variant effect prediction using deep learning-derived splice scores	1	278	2	
Linder J, 2022, Nat Mach Intell	Interpreting neural networks for biological sequences by learning stochastic masks	1	6	2	
Armenteros JJA, 2017, Bioinformatics	Deeploc: prediction of protein subcellular localization using deep learning	4	637	3	Green
Gronning AGB, 2020, Nucleic Acids Pic	Deepclip: predicting the effect of mutations on protein-rna binding with deep learning	3	47	3	
Koo PK, 2021, Plos Comput Biol	Global importance analysis: an interpretability method to quantify importance of genomic features in deep neural networks	1	23	3	
Gray, 2018, Cell Syst	Quantitative missense varıant effect prediction using large-scale mutagenesis data	7	113	4	Purple
Frazer J, 2021, Nature	Dısease varıant prediction with deep generative models of evolutionary data	9	156	4	
Marquet C, 2022, Hum Genet	Embeddings from protein language models prediction conservation and variable effects	3	33	4	
Dunham AS, 2023, Genome Biol	High-throughput deep learning variable effect prediction with sequence unet	1	5	4	
Jagota M, 2023, Genome Biol	Cross-protein transfer learning substantıally improves disease variable prediction	1	4	4	
Heyne HO, 2020, Sci Transl Med	Predicting functional effects of missense variants in voltage-gated sodium and calcium channels	4	60	5	Orange

Bosselmann CM, 2022, Ebiomedicine	Predicting the functional effects of voltage-gated potassium channel missense variants with multi-task learning	1	5	5	
Siedhoff NE, 2021, J Chem Inf Model	Pypef-an integrated framework for data-driven protein engineering	2	13	6	Brown
Wittmund M, 2022, Acs Catal	Learning epistasis and residual coevolution patterns: current trends and future perspectives for advancing enzyme engineering	2	17	6	

3.5. Keywords

In this section, a keyword analysis was conducted on publications pertaining to the utilization of machine learning methodologies in variant effects prediction. The most salient keywords from the publications are presented as a word cloud in Figure 13 and a tree map according to the rates in Figure 14.

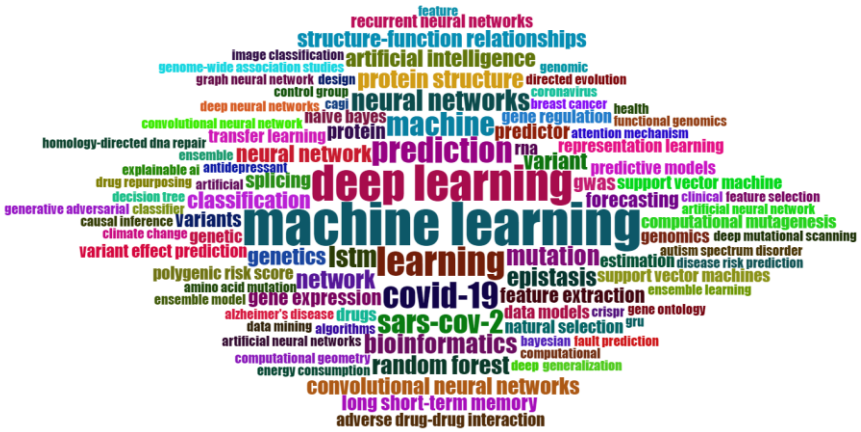

Figure 13. WordCloud

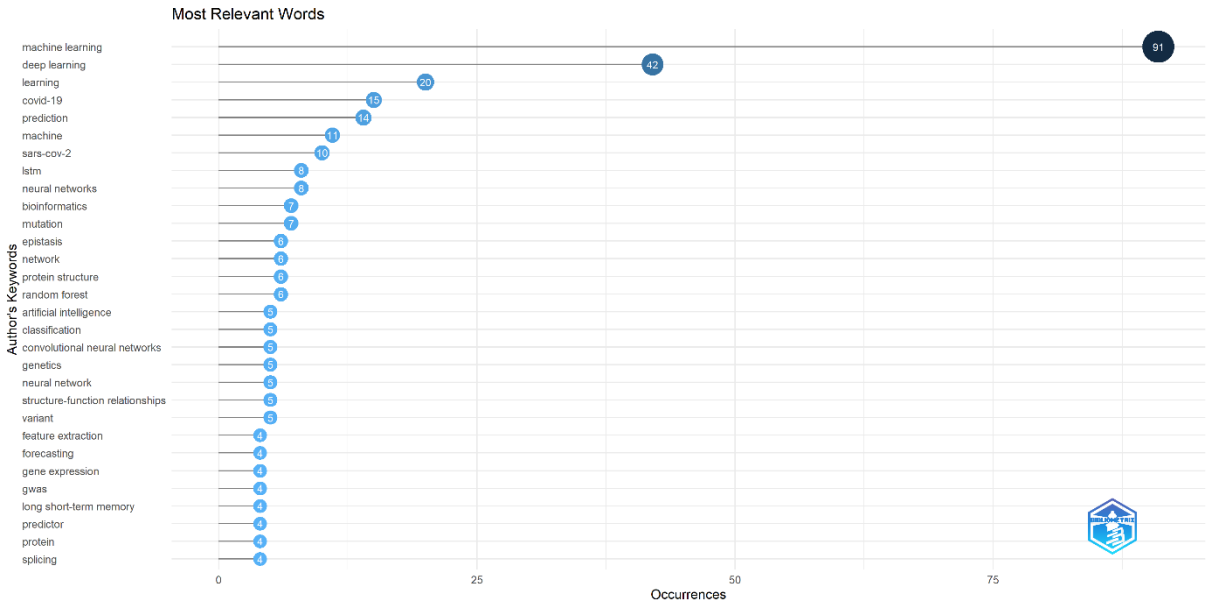

Figure 14. The frequencies of the 30 most frequently used keywords.

Figure 13 presents a visual representation of the 100 most frequently used keywords, as identified through the word cloud analysis. Figure 14 presents a graphical representation of the frequencies of the

30 most frequently used keywords. The most frequently used keywords in publications on the use of machine learning methods in variant effect prediction, as identified in the analyses presented in Figures 13 and 14, were "machine learning", "deep learning", "learning", "covid-19", "prediction", "machine" and "sars-cov-2". Figure 15 illustrates the cumulative distribution of keywords by year.

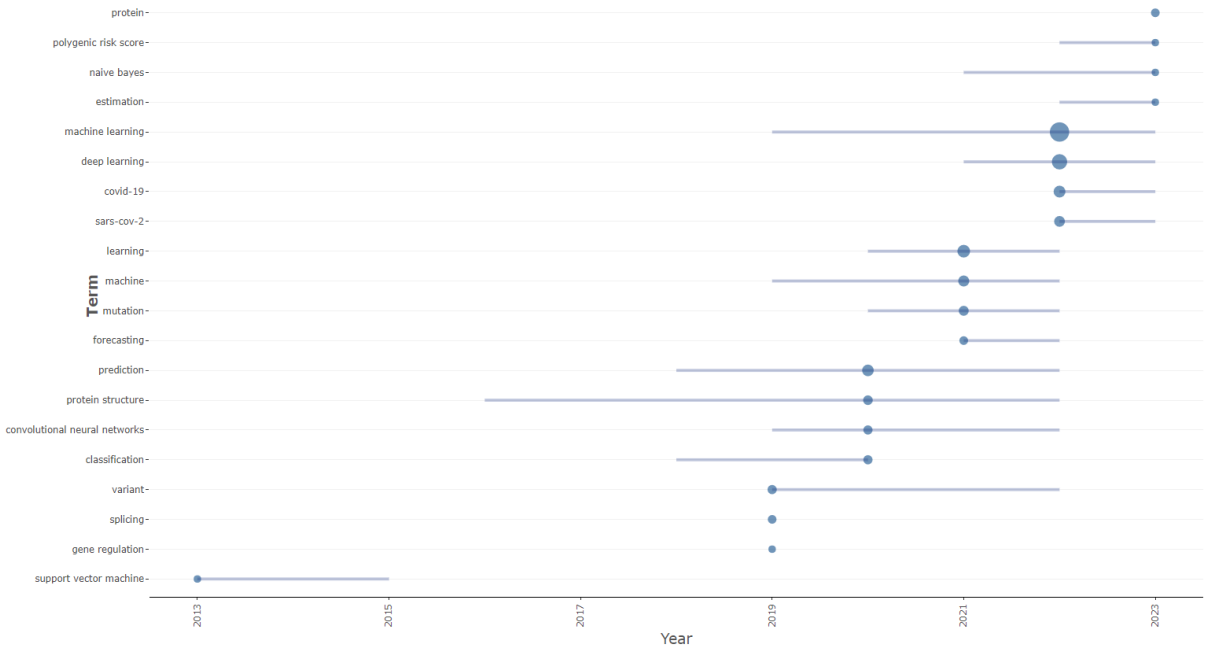

Figure 15. The cumulative distribution of keywords by year

Upon examination of the graph provided in Figure 15, it becomes evident that the most frequently utilized keywords over time are "protein structure", "prediction", "machine learning", "variant", "convolutional neural network", and "machine". In recent years, there has been a notable increase in the usage of specific keywords, including "deep learning", "naive bayes", "covid-19", "sars-cov-2," and "polygenic risk score". Thematic analysis of publications on the utilization of machine learning methodologies in variant effect prediction was conducted periodically, with the relationships between identified themes illustrated in Figure 16.

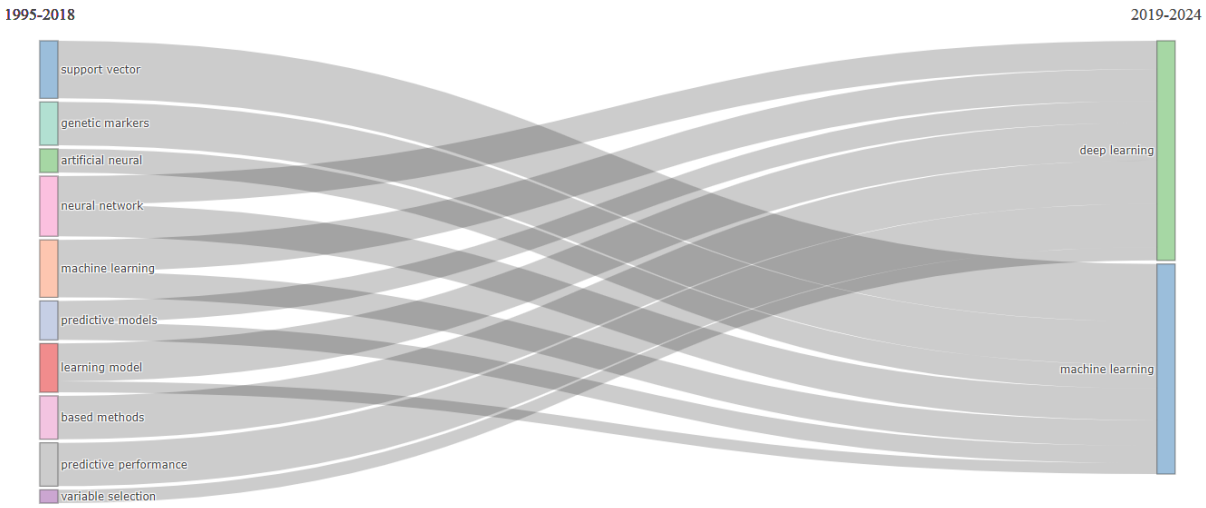

Figure 16. Thematic change

As illustrated in Figure 16, the themes between 1995 and 2018 were “support vector”, “genetic markers”, “artificial neural”, “neural network”, “machine learning”, “predictive models”, “learning model”, “based methods”, “predictive performance”, “variable selection”, and between 2019 and 2024 the themes were “deep learning” and “machine learning”.

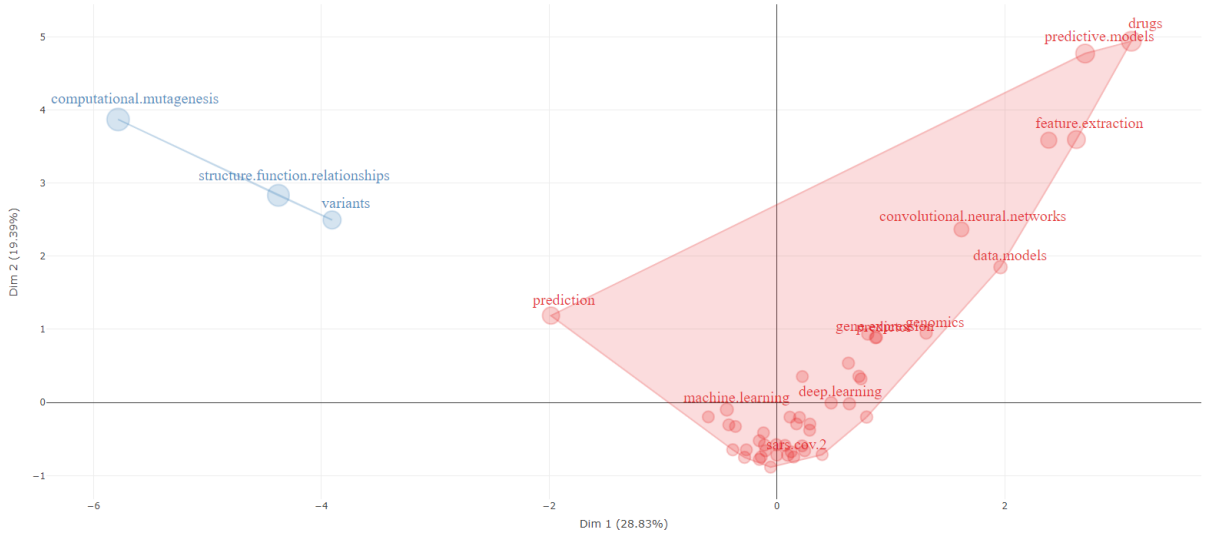

Figure 17. Factor analysis of the publications

The word pairs of keywords used together in the publications were subjected to factor analysis (Figure 17). Accordingly, the red cluster was found to have been researched on several topics, including machine learning, deep learning, Covid-19, Sars-cov-2, long short-term memory (Lstm), neural networks, bioinformatics, protein structure, and epistasis. In contrast, the blue cluster was found to have been researched on some topics, including computational mutagenesis, structure-function relationship, and

the variants. The keyword co-occurrence network employed in the aforementioned publications is presented in Figure 18.

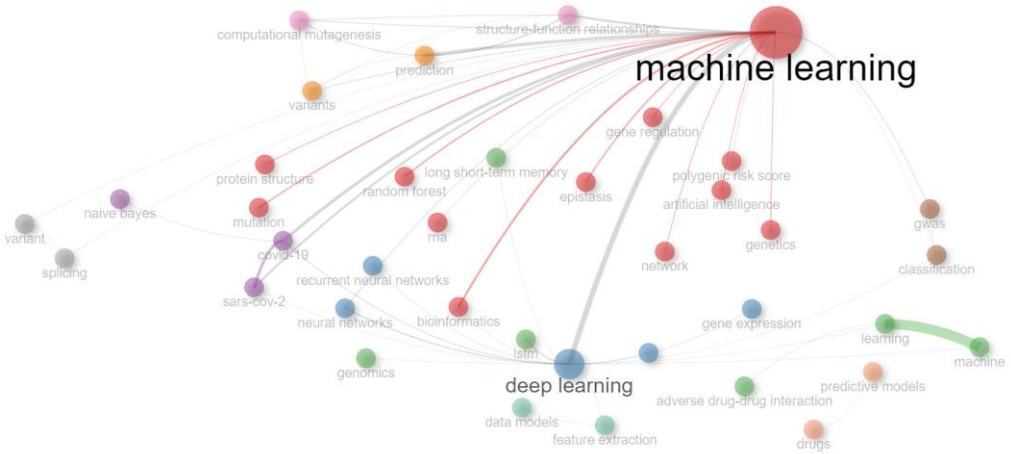

Figure 18. The keyword co-occurrence network of the publications

Upon examination of the keyword co-occurrence network presented in Figure 18, it becomes evident that the keywords utilized in publications on the application of machine learning methodologies in variant effect prediction predominantly constitute a keyword cluster with the concepts of "machine learning" and "deep learning." Moreover, the keywords most frequently occurring in conjunction with other keywords employed in these publications are also discernible within the network. Figure 19 presents a network analysis illustrating the three-field relationship between the keywords employed by the authors and the journals.

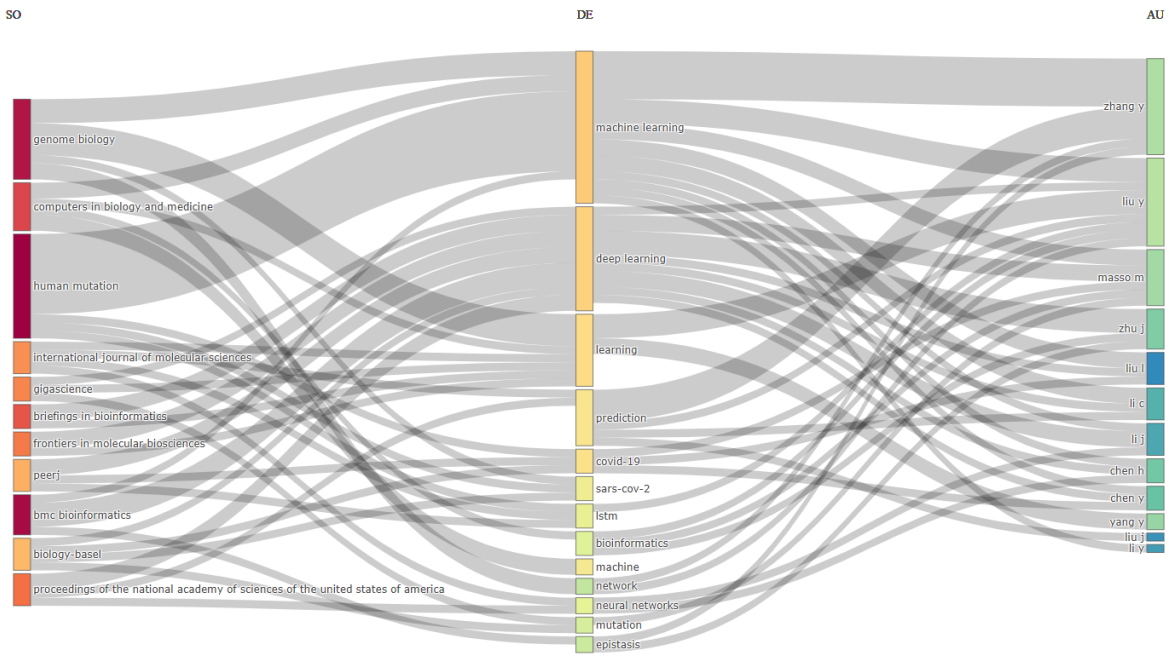

Figure 19. Three-field plot for journals, authors and keywords

Figure 19 illustrates the distribution of journals on the left, frequently used keywords in the middle, and authors on the right. It was observed that the most frequently used keywords, "machine learning" and "deep learning" were utilized by nearly all authors and journals represented in the graph. To gain a comprehensive understanding of the keywords employed in the publications, the abstract sections were also examined, and topics were identified (Figure 20).

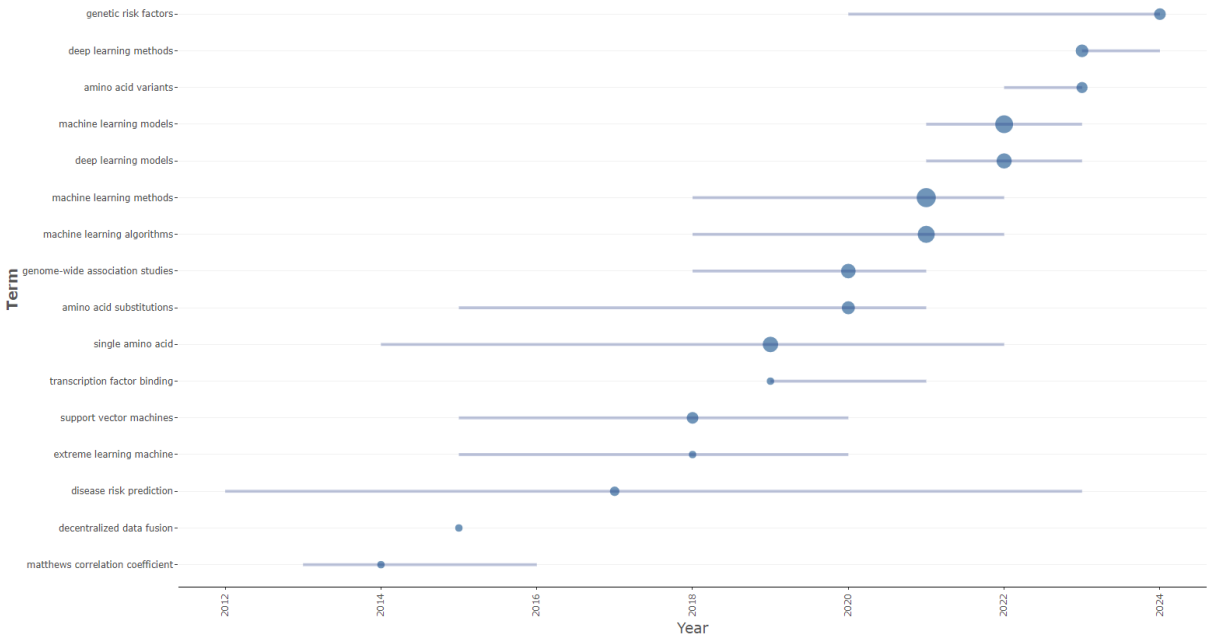

Figure 18. Trend topic for keywords used in the abstract section

In the graph presented in Figure 18, it was found that the trend topic in 2017 was “disease risk prediction”, in 2018 “support vector machine”, in 2019 “single amino acid”, in 2020 “amino acid substitutions” and “genome-wide association studies”, in 2021 “machine learning methods and algorithms”, in 2022 “deep learning” and “machine learning models”, in 2023 “amino acid variant” and “deep learning methods”, and in 2024 “genetic risk factors”.

4. Conclusion, Discussion and Limitations

While a comprehensive analysis of the utilisation of machine learning methodologies for the prediction of variant effect between 1995 and 2024 (up to March 2024) has been undertaken, this study is not without its limitations: (i) The research is limited to a time period from 1995 to March 2024. (ii) The database utilized in this study, which serves as a valuable resource for examining the application of machine learning techniques in variant effect prediction, is limited to WoSCC. The search conducted in PubMed and Scopus databases returned a smaller number of publications than the search conducted in WoSCC. However, in future studies, more comprehensive investigations can be conducted using databases such as Google Scholar, Science Direct, and Elsevier. (iii) This study included only research articles and review articles; other scientific publications were not analyzed. In future studies, the scope

can be expanded by including different scientific publications, such as book chapters, theses, papers, preprints, comments, and letters to the editor.

In this study, a bibliometric method was employed to characterise the scientific research conducted between 1995 and 2024 on the utilization of machine learning methods in variant effect prediction. This approach was undertaken to ascertain the status and focus of research on this subject, with a comprehensive analysis of the studies in question. Consequently, it has been observed that the number of publications in this field has increased at a steady rate over the past eight years. This growth demonstrates that machine learning methods have been increasingly applied in variant effect prediction in recent years. Similarly, the frequency of citations indicates that studies investigating the use of machine learning methods in variant effect prediction are rapidly increasing.

A review of the distribution of publications by country reveals that the United States has the highest total publication volume and the highest total citation volume. It can thus be concluded that the United States plays a pivotal role in the research on the utilisation of machine learning techniques in the domain of variant effect prediction, having made substantial contributions to this field. A review of international collaboration reveals that the United States engages in joint research with China, Germany, England, and Australia. In terms of publishing institutions, it is evident that eight of the top ten most productive institutions are located in the United States, with the remaining two situated in Germany. An analysis of the citation numbers of countries revealed that institutions in the United States had significantly higher citations than other countries. Consequently, it can be concluded that the United States has a prominent and extensive presence in this field, characterised by a substantial publication volume, a considerable number of citations, and a notable level of international collaboration.

A review of the literature revealed that the majority of publications on the use of machine learning methods in variant effect prediction were found in the Human Mutation journal, while the Nucleic Acids Research journal was the most frequently cited. An analysis of the authors' citation levels according to publication year and journal of publication reveals that the highest number of citations were made to an article by Zhou and Troyanskaya (2015), published in Nature Methods. A LCS and GLC analysis of between-author citations revealed the existence of a citation network comprising six clusters. In this clustering based on subject areas, the publication with the highest LCS value was that of Frazer et al. (2021), published in Nature. The highest GLC value was observed in the publication by Armenteros et al. (2017), in Bioinformatics. The aforementioned authors and their publications have made a substantial contribution to the field of machine learning-based variant effect prediction and are included in the list of the ten most productive authors and the ten most cited authors. It would be beneficial for researchers planning to study the use of machine learning methods in variant effect prediction to examine these authors and publications. Furthermore, it is important to be aware of these publications, as they have made significant contributions to the field.

According to the keyword analysis for publications on the use of machine learning methods in variant effect prediction, the prominent keywords in the publications are: "machine learning", "deep learning".

A review of the thematic evolution of the studies revealed that the themes of support vector, genetic markers, artificial neural network, neural network, machine learning, predictive models, learning models, based methods, predictive performance, and variable selection were examined between 1995 and 2018. After 2019, it was determined that the thematic focus shifted to machine learning and deep learning. It can be posited that the rationale behind this shift in focus is the advent of novel deep learning models, which have begun to emerge alongside traditional machine learning methods in the domain of variant effect prediction studies (Qi et al., 2021; Rentzsch et al., 2021). In studies employing traditional variant effect prediction models, the necessary information for determining the effect of a variant on the phenotype with the prediction model is evolutionary and structural data about the variant (Riesselman et al., 2018). However, in studies developing deep learning-based prediction models, inferences can be made about the variant using the raw data of the variant (Jiang et al., 2021). Therefore, it can be stated that machine learning and deep learning themes have gained importance in studies conducted since 2019.

A review of the abstracts of published works reveals that, in addition to machine learning and deep learning, the subjects most frequently investigated by researchers are "disease risk prediction," "single amino acid", "amino acid substitutions", "genome-wide association studies", "amino acid variant", and "genetic risk factors." It is anticipated that these identified topics will inform the direction of future research in this field. Furthermore, an analysis of the keywords used by the authors according to journals was also provided. Our findings demonstrate that the publications were shaped according to the concept of machine learning and deep learning. Machine learning, deep learning, Covid-19, Sars-cov-2, long short-term memory (Lstm), neural networks, bioinformatics, protein structure, and epistasis form a cluster; computational mutagenesis, structure function relationship, and the variants form a cluster and are interconnected. It can be understood from this that amino acid and genome studies are of great importance in variant effect prediction. Furthermore, the current pandemic has highlighted the necessity for variant effect prediction in this area. Finally, structural information of the variant is also an important research topic in variant effect prediction.

This study presents an overview of the various themes and publications that have emerged over time on the use of machine learning methods in variant effect prediction. This thematic shift has revealed that the most extensively researched areas by researchers, academics and universities are machine learning and deep learning. The topics investigated in this field include amino acid changes, mutations at the genome level, structural information of variants, covid-19 mutations and protein structure. In future studies, these research topics can be investigated with different methods based on machine learning and deep learning.

Note

This article was presented as an oral presentation at the 15th Medical Informatics Congress held in Trabzon on 30-31 May 2024.

Conflict of interest: No conflict of interest has been declared by the authors.

Authorship contributions: The contribution of the authors is equal.

References

Almagro Armenteros JJ., Sønderby CK., Sønderby SK., Nielsen H., Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 2017; 33(21): 3387-3395.

Angermueller C., Pärnamaa T., Parts L., Stegle O. Deep learning for computational biology. Molecular Systems Biology 2016; 12(7): 878-894.

Aria M., Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics 2017; 11(4): 959-975.

Bromberg Y., Prabakaran R., Kabir A., Shehu A. Variant effect prediction in the age of machine learning. Cold Spring Harbor Perspectives in Biology 2024; 16(7): a041467.

Donthu N., Kumar S., Mukherjee D., Pandey N., Lim WM. How to conduct a bibliometric analysis: An overview and guidelines. J. Bus. Res. 2021; 133: 285-296.

Fidanoğlu P., Belder N., Erdoğan B., İlk Ö., Rajabli F., Özdağ H. Genom projeleri 5N1H: Ne, nerede, ne zaman, nasıl, neden ve hangi popülasyonda? Türk Hijyen ve Deneysel Biyoloji Dergisi 2013; 71(1): 45-60.

Frazer J., Notin P., Dias M., Gomez A., Min JK., Brock K., Gal Y., Marks DS. Disease variant prediction with deep generative models of evolutionary data. Nature 2021; 599(7883): 91-95.

Horne J., Shukla D. Recent advances in machine learning variant effect prediction tools for protein engineering. Industrial and Engineering Chemistry Research 2022; 61(19): 6235-6245.

Ionita-Laza I., Mccallum K., Xu B., Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nature Genetics 2016; 48(2): 214–220.

Jiang T., Fang L., Wang K. Deciphering the language of nature: A transformer-based language model for deleterious mutations in proteins. The Innovation 2021; 4(5).

Li MX., Gui HS., Kwan JSH., Bao S.Y, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Research 2012; 40(7): e53.

Livesey BJ., Marsh JA. Advancing variant effect prediction using protein language models. Nature Genetics 2023; 55(9): 1426-1427.

Mahmood K, Jung CH., Philip G., Georgeson P., Chung J., Pope BJ., Park DJ. Variant effect prediction tools assessed using independent, functional assay-based datasets: Implications for discovery and diagnostics. Human Genomics 2017; 11: 1–8.

Niroula A., Vihinen M. Variation interpretation predictors: Principles, types, performance, and choice. Human Mutation 2016; 37(6): 579–597.

Niroula A., Vihinen M. How good are pathogenicity predictors in detecting benign variants? PLoS Computational Biology 2019; 15: 1–17.

Qi H., Zhang H., Zhao Y., Chen C., Long JJ., Chung WK., Guan Y., Shen Y. MVP predicts the pathogenicity of missense variants by deep learning. Nature Communications 2021; 12(1): 510.

Qiu J., Nechaev D., Rost B. Protein-protein and protein-nucleic acid binding residues are important for common and rare sequence variants in human. BMC Bioinformatics 2020; 21: 452.

Qu H., Fang X. A brief review on the human encyclopedia of DNA elements (ENCODE) project. Genomics Proteomics Bioinformatics 2013; 11(3): 135–141.

Rentzsch P., Schubach M., Shendure J., Kircher M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Medicine 2021; 13: 1-12.

Riesselman AJ., Ingraham JB., Marks DS. Deep generative models of genetic variation capture the effects of mutations. Nature Methods 2018; 15(10): 816-822.

Tang H., Thomas PD. Tools for predicting the functional impact of nonsynonymous genetic variation. Genetics 2016; 203(2): 635–647.

The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007; 447: 799-816.

The International HapMap Consortium. The international HapMap project. Nature 2003; 426: 789-796.

The 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. Nature 2010; 467: 1061-1073.

Xu F., Guo G., Zhu F., Tan X., Fan L. Protein deep profile and model predictions for identifying the causal genes of male infertility based on deep learning. Information Fusion 2021; 75: 70-89.

Zhou J., Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods 2015; 12(10): 931-934.