Journal of Intellectual Property Rights
Vol 30, July 2025, pp 481-488
DOI: 10.56042/jipr.v30i4.12447

Use of Machine Learning at the Patent office to Track Global Trends in Healthcare Innovation

Ashwini Siwal† and Chinmay

Faculty of Law, University of Delhi, Delhi — 110 007, India

Received: 13th July 2024; revised: 7th May 2025

The present day world comprises several complex systems. These intricate systems require complex and sophisticated tools for their analysis in order to make precise forecasts about them which might enhance societal well-being. Use of modern day disruptive technologies based on artificial intelligence can be used to comprehend many of these intricate systems which remain unexplored by professionals. Accurate machine-learning algorithms have potential use in several disciplines such as medicine, climate change, traffic patterns, and criminal recidivism. This paper endeavours to leverage machine learning for analysis of patent data in order to facilitate innovation in the public health sector. The research focussed on the objective of developing computational methods to classify and cluster patent documents. Thus, aiding policy makers, legal experts and the research community in navigating the rapidly evolving landscape of patented healthcare technology. A three-phase approach methodology was adopted to meet the objective of the research. The first phase comprised data collection and processing, the second phase focussed on machine learning model development and the final phase included application and validation of the developed tools. The study became relevant in the wake of the fourth industrial revolution and the Covid Pandemic which has resulted in enhancing the significance of the healthcare industry exponentially. This exponential growth has resurfaced certain pertinent issues at the interface of intellectual property regime and the public healthcare system. Particularly the implications of the patent system on public health are highly debated. It becomes very crucial, more so in the case of a developing nation, to study the effect of the patent system on innovation, technology transfer and industrial dynamics with reference to the pharmaceutical sector. This study will be crucial in addressing some of these issues which concludes with development of ML based patent analytics tool and suggests that such an investment in machine learning based patent analytics tool helps to reduce patent prosecution and litigation thus providing a cost effective solution.

Keywords: Patent Grants, Machine Learning, Public Health, Healthcare Innovation, Technology Transfer, Patent Analytics, Evidence-based Policy Discussions

Sophisticated research tools are valuable for analysing intricate systems. Health Innovation is only one of many such intricate systems. There are several processes in the universe that are not yet comprehended by professionals, and making precise forecasts about them might enhance societal well-being. Accurate machine-learning algorithms have potential use in several disciplines such as medicine, climate change, traffic patterns, and criminal recidivism. Healthcare Innovation is a field characterised by intricate systems that human specialists frequently struggle to comprehend. In such cases, machine learning can provide valuable assistance.

Artificial intelligence (AI) based machine learning tools have made significant progress in the field of healthcare. It has the capacity to deliver improved therapeutic outcomes at a lower cost. Simultaneously, however, it has been demonstrated that AI models relying on inappropriate data can make clinically incorrect choices, often targeting historically marginalised populations with these flaws in a systematic manner. In order to prevent unforeseen negative consequences, participants in the development and adoption process must actively encourage and uphold responsibility and authenticity.

There is immense scientific and public discussion on the employment of predictive algorithms by government actors. Conventional linear regression is a data model created by humans that simply converts inputs into outputs. Alternatively, the decision rule might be derived via algorithmic or machine learning techniques also. Machine learning involves a range of algorithms with different levels of complexity. One key characteristic of this field is that the learning algorithm does not explicitly define the decision rule. Instead, it "learns" the decision rules using training data. Commentary in both instances has frequently

†Corresponding author: Email: asiwal@law.du.ac.in

been quite critical, especially when discussing the implementation of algorithms in fields such as predictive policing and criminal risk assessment.

This paper endeavours to leverage machine learning for analysis of patent data in order to facilitate innovation in the public health sector. The research focussed on the objective of developing computational methods to classify and cluster patent documents. Thus aiding policy makers, legal experts and the research community in navigating the rapidly evolving landscape of patented healthcare technology. The following sections give a background for the study by providing the reader about the origins of bioinformatics as to how disruptive technologies embarked in the arena of patent healthcare innovation. Further the paper discusses the working and importance of machine learning; quality of patents and how to assess the quality of the patents. After setting the background the paper in the next section provides the methodology of the research and finally provides the findings before concluding the research.

The Origin of Bioinformatics

In 1999, the effort to sequence the human genome was working at full capacity, nine years after the National Institutes of Health's National Centre for Human Genome Research had announced its initial collaborative research plan. The complete genetic material of an independent creature, Haemophilus influenzae, had already been decoded, and the initial comprehensive sequencing of a human chromosome will be made public in the same year. The next year, President Bill Clinton and United Kingdom Prime Minister Tony Blair would jointly reveal the preliminary version of the human genome, commonly referred to as the rough draft. When it comes to bioinformatics, we have the advantage of having a historical record. The organised gathering and examination of biological sequencing data has required the combined expertise of computer scientists, statisticians, and biologists for more than forty years, with the word "bioinformatics" being coined in 1970.[1] The concept of a bioinformatician, as we understand it now, emerged in the late 1990s and early 2000s. In addition, colleges such as the University of Michigan and the University of California San Diego have started to offer training and degree programmes in bioinformatics.

Unsurprisingly, the Patent Offices around the world have observed an increasing influx of patent applications for bioinformatics technologies. According to comments from the sector, it is anticipated that there would be a significant increase in the following years. In December of 1999, the USPTO had created a new art section specifically dedicated to the examination of bioinformatics applications in a standardised manner. The art unit is located at USPTO Technology Centre 1600 and specialises in the examination of inventions in the fields of biotechnology and organic chemistry. It is specifically designated as art unit 1631. The patent examiners in AU 1631 have a wide range of knowledge and skills, encompassing not only the field of biological sciences, but also physics, electrical engineering, and, notably, computer science. The USPTO deemed the software and data processing patent cases from the late 1990s to be directly applicable to patents on computing tools used for analysing biological systems. As a result, the USPTO recommended that bioinformatics inventors take guidance from the software invention guidelines outlined in the Manual of Patent Examining Procedure ("MPEP").The user's text is empty. Consequently, the significant increase in the capacity to patent software around that time had a direct impact on AU 1631.

How Machine Learning Works?

Machine learning algorithms acquire knowledge directly from the data. In the context of "supervised learning," which is the main emphasis of this study, the data scientist provides the learning algorithm with data that has been carefully selected and labelled by experts in the area, based on input characteristics and output labels. In order to understand the working of machine learning better an algorithm may be thought to have been fed 100,000 X-rays of human lungs, out of which 5,000 have been identified by radiologists as displaying malignant tumours. If the radiologists and data scientists have performed their tasks accurately, this dataset should approximate the "ground truth," which refers to the true and accurate underlying reality.

The process of learning or "training" a model entails iteratively adjusting the model parameters to optimise the translation of inputs into outputs.[1]The model's prediction performance is continuously evaluated and improved until no more improvements can be made. After the machine learning model has undergone training, it is often checked on a portion of the training data that it has not seen before, referred to as "test" data, in order to evaluate its performance on

new data. Ideally, the model is also tested on data from a completely other source to further confirm its effectiveness.

The Advantages of Using Machine Learning

Examiners perform prior art searches by utilising technology classes and keywords to search databases of the Patent Office and other internet sources. Keyword search is not very useful for software-related applications due to the inconsistent wording used to explain the same topic. Despite relying on keyword search, examiners are aware of its limitations. Computer scientists at the Patent Office have expressed criticism against keyword search, highlighting that "basic keyword searches have restricted effectiveness in the patent prosecution context".[2] It is particularly advantageous to apply this approach to a multitude of systems (whether they are physical, social, or a combination of both) that have not yet been comprehended by human area specialists. Machine learning may be highly beneficial in situations where it can analyse several input factors that may be important and, if well verified and validated, can produce precise predictions.

The Conundrum Surrounding the Quality of the Patent

A significant portion of the present litigation arises from apprehensions over the strength of patents. Many analysts have suggested plans for enhancing quality. Critics express worry about several topics when they lament poor quality. These include: non-compliance with the statutory requirement that patents be granted only to inventions that are not obvious to the average scientist or technologist in the field; non-compliance with the statutory requirement that the patent provide instructions on how to fully make and use the invention described in the claims; and violation of the principle that the patent claims must clearly indicate the boundaries of the patent rights.[3] The shortcomings of patent quality are most evident in legal disputes, primarily because patent litigation is associated with exorbitant expenses, sometimes unpredictable results, and the potential for long-term negative impacts on innovation. Instead of being the source of low patent quality, one significant area of suggested improvements is the infrastructure for the initial assessment of patents.

Evaluating the Standard of Patents

The task of characterising patent quality using qualitative words is equally difficult as the task of describing and evaluating patent quality using quantitative criteria. The challenge stems from the fact that the existing research has not consistently differentiated between three separate concepts of quality: (1) a patent document that is considered "important" and aids in the spread of knowledge; (2) the personal value of a patent to its owner; and (3) a patent's adherence to the established legal requirements for patentability.[4] Furthermore, a growing body of research examines the concept of quality by considering the qualities of examiners and the incentives that influence the thoroughness of the examination process.

Quantitative Metrics for Assessing the Validity of Patents

Some of the important metrics to assess validity of patents include Citation, knowledge transfer and technology transfer; Private value of patent; Legal Validity; Characteristics of Examiner. The literature documenting the forward citations of patents emphasises the patent's function in disseminating scientific or technical information within a certain group. According to this perspective, patents that receive a significant number of citations from other patents are likely to be of great significance. The number of co-inventors is strongly connected with forward citation rates and technological relevance. Further empirical studies have demonstrated that examiners exert a substantial effect on citations. Researchers frequently use proxies in the literature to estimate private value. One common approach is to examine the features of patents that are involved in litigation or for which renewal fees are paid. Both scenarios assume that sensible entities would not spend money without expecting a financial gain in return. Research in this field has established that patents that have been challenged or renewed tend to have a greater number of claims and are also more likely to be cited by other patents in both forward and backward directions.[5]Therefore, some analysts have utilised the quantity and intricacy of claims not just as an indication of the patent's worth, but also as a measure of the work put into getting it.

A further significant factor that is evidently linked to the value of private patents is the magnitude of the patent family, which refers to the number of foreign jurisdictions where the applicant has simultaneously sought patent protection for the identical innovation. A growing corpus of literature investigates the interplay between characteristics of examiners and

their impact on the quality of patents. The function of technology, as well as the involvement of examiners in patent quality, is intricate. Given that certain components of an examiner's job need expertise in technology, it is essential for the examiner to possess training in a relevant scientific or technical area.

In addition to subject matter, institutional incentives also contribute to significant heterogeneity in examiner behaviour, which has significant consequences for patent quality. Experience level may impact the quality of work, and this relationship can be either favourable or negative. Veteran examiners may provide more knowledgeable assessments, leading to a positive connection. On the other hand, amateur examiners may pay more attention to details, resulting in a negative association. Experience also impacts the amount of time provided to examiners. The presence of time pressure might intensify any possible feelings of "burnout" experienced by veterans.The quality of patents, especially in terms of notice, may be influenced by a combination of examiner training and a high level of knowledge.[6] Therefore, training examiners, especially in the application of the written description requirement, and in methods to compel patent applicants to clarify the interpretation of potentially troublesome claim terms, are expected to have positive results. In addition, providing training to examiners on fundamental concepts of scientific peer review might be beneficial.

Computer science had previously developed tools that were also used in the biological sciences field, namely in bioinformatics as noted above; however these technologies had not yet been organised into a distinct discipline. The development of specialised tools to handle the unique breadth and magnitude of patent information heralded the inception as a distinct field of study.[7]This research is a step in that regard. Further as noted above, the domain of healthcare innovation concerning patents requires measures to improve inter alia infrastructure for assessment of patents, evaluating the standard of patents and assessing the quality by considering important patent documents. This necessitates ML based analytical techniques to be used effectively for addressing these needs. This study is an attempt to leverage machine learning to analyse patent data to facilitate innovation in the public health sector.

Objective of the Study

The overall objective of the study is to leverage machine learning techniques to analyse patent data in order to facilitate innovation in the public health sector. The study focuses on developing computational methods to classify and cluster patent documents in order to aid policymakers, legal experts, and the research community in navigating the rapidly evolving landscape of patented healthcare technology. The study was carried forward by adopting a three-phase approach methodology overarching three objectives. The first objective was to investigate the Role of Patents in Healthcare Innovation. The second objective was to develop Intelligent Analytics for Patent Data. The third objective consisted of Support Continuous Tracking of Healthcare Innovations. The detailed methodology for each objective is discussed in the next section.

Methodology

In order to achieve the objective of leveraging machine learning to analyse patent data, facilitating innovation in the public health sector, the study focused on developing computational methods to classify and cluster patent documents in order to aid policymakers, legal experts, and the research community in navigating the rapidly evolving landscape of patented healthcare technology. The methodology adopted was a three-phase approach comprising data collection and preprocessing, machine learning model development, and application and validation of the developed tools. Each phase of the methodology helped to achieve one objective. The first objective was to investigate the Role of Patents in Healthcare Innovation. In order to fulfil this objective patent data was analysed to identify the impact of patents on public health innovation. Machine learning models were utilised to extract and classify information from patent documents. The second objective was to develop Intelligent Analytics for Patent Data which was achieved by implementing BERT and LDA models for text representation and topic generation. K-means clustering was used for identifying technology clusters.[8] The third objective consisted of Support Continuous Tracking of Healthcare Innovations which was achieved by designing and testing a tool to monitor healthcare technology developments post-National IPR Policy 2016. The tool was deployed for policy analysis and educational purposes.

Findings under Objective 1

To Investigate the Role of Patents in Healthcare Innovation

The investigation into the role of patents in healthcare innovation began with an extensive data collection process. The sources of data included major patent offices such as the Office of

the Controller General of Patents, Designs &TradeMarks (CGPDTM) in India, the United States Patent and Trademark Office (USPTO), the State Intellectual Property Office of China (SIPO), and the World Intellectual Property Organization (WIPO). Data spanning from 2015 to 2021 was collected, amounting to over 134,728 patent documents relevant to public health innovations. The collected data was pre-processed to ensure consistency and usability. This involved cleaning the data to remove any irrelevant information, normalising the formats, and organising the data into a structured database. The data preprocessing was performed by a team of four MSc students under the supervision of Dr. Usharani Hareesh Govindarajan. These students were responsible for writing scripts to automate the data-cleaning process, which included removing duplicates, correcting errors, and standardising the terminology used in the patent documents. Following the preprocessing, the data was analysed using a machine-learning approach.[9] The project employed Bidirectional Encoder Representations from Transformers (BERT) for text representation and Linear Discriminant Analysis (LDA) for topic generation. BERT was chosen for its ability to understand the context of words in patent documents, making it suitable for extracting meaningful information. The LDA model was used to generate topics from the patent texts, helping to identify the main areas of innovation. This phase of the project was led by Dr. Ashwini Siwal, who provided the legal perspective and was executed by a team of two PhD students specialising in machine learning.

Results

The analysis revealed key technology trends and the convergence of multiple technologies in the public health sector. Technologies like blockchain, AI, and machine learning are poised to transform healthcare but are hindered by patent barriers. The research developed a machine learning framework to analyse patents, fostering a better understanding of innovation in public health. The insights gained from this analysis were used to identify patents that have the potential to drive significant advancements in healthcare. These results were discussed in bi-weekly meetings with the entire project team, including both legal and technical experts, to ensure a comprehensive understanding and validation of the findings.

Findings under Objective 2

To Develop Intelligent Analytics for Patent Data

Model development the development of intelligent analytics for patent data involved the implementation of advanced machine learning models. The project utilised Bidirectional Encoder Representations from Transformers (BERT) for text representation with Latent Dirichlet Allocation (LDA) for topic modelling. The BERT model (Fig. 1) was fine-tuned on the collected patent dataset to adapt to the specific terminology and context of healthcare patents. These extracted classifications are further followed through Kmeans clustering that enables the identification of

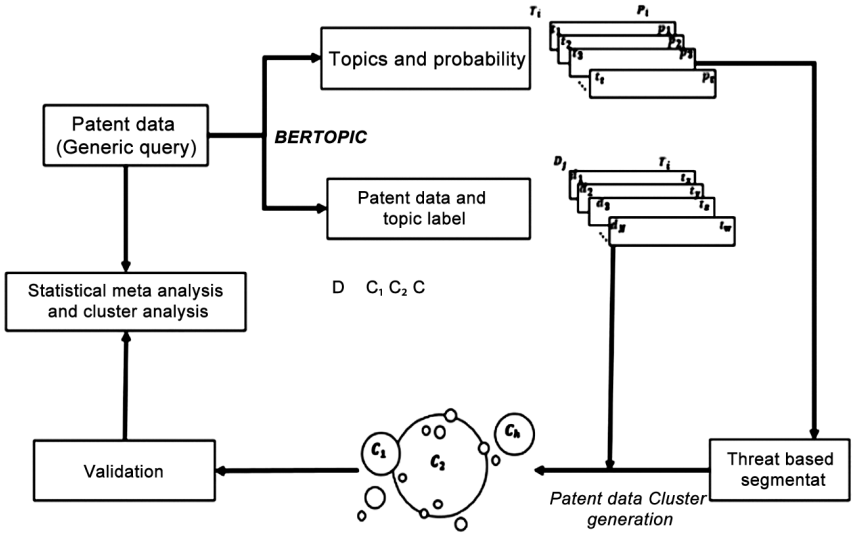

Fig. 1 — Computation flow diagram

key technology clusters followed by objective validations. This aids intelligent analytics of large-sized patent grant datasets on the healthcare topic, through the generated vector information, for clustering the patents and aiding in rapid legal/natural for effective knowledge management. The analytics insights result of the machine learning model accelerates patented healthcare innovation.

Model Training Details

- Instances: The model was trained on 134,728 patent documents.
- Training Time: Training was conducted over 48 hours on a high-performance computing cluster with 64 cores and 256GB of RAM.
- Evaluation: The topic coherence score varied with different minimum cluster sizes, achieving the highest coherence score of 0.610 at a minimum cluster size of 34.

Clustering

Post classification, K-means clustering was applied to group the patents into technology clusters. This helped in identifying major areas of innovation and the relationships between different technological advancements. The clustering process involved several iterations to determine the optimal number of clusters, which was found to be 15. Validation of these clusters was performed using the Stepwise Weight Assessment Ratio Analysis (SWARA) technique to ensure the robustness of the clusters. The development and fine-tuning of these models were carried out by a team of four MSc students in computer science, who worked under the guidance of Dr. Usharani Hareesh Govindarajan. The students developed the initial codebase, performed hyper-parameter tuning, and conducted the evaluations to ensure the models met the desired performance metrics. The code base will be openly published in Github for further enhancements Fig. 2.

Results

The research methodology employing machine learning to analyze patent data aims to uncover key technology trends and clusters within healthcare patents, streamline knowledge management, and accelerate innovation. It provides policy and legal support by offering insights into the impact of the National IPR Policy 2016 and subsequent amendments. A comprehensive quantitative analysis of contemporary research aligns with public health needs, assessing the relevance of current technologies in public health infrastructure. The methodology also supports stakeholders in decision-making, aids the research community in selecting advanced patented technologies, and has an educational and social impact by integrating the analysis framework into academic networks. This facilitates technology transfer and ensures healthcare systems are updated with the latest technological advancements, ultimately fostering a dynamic and innovative public health sector. The insights included the identification of key

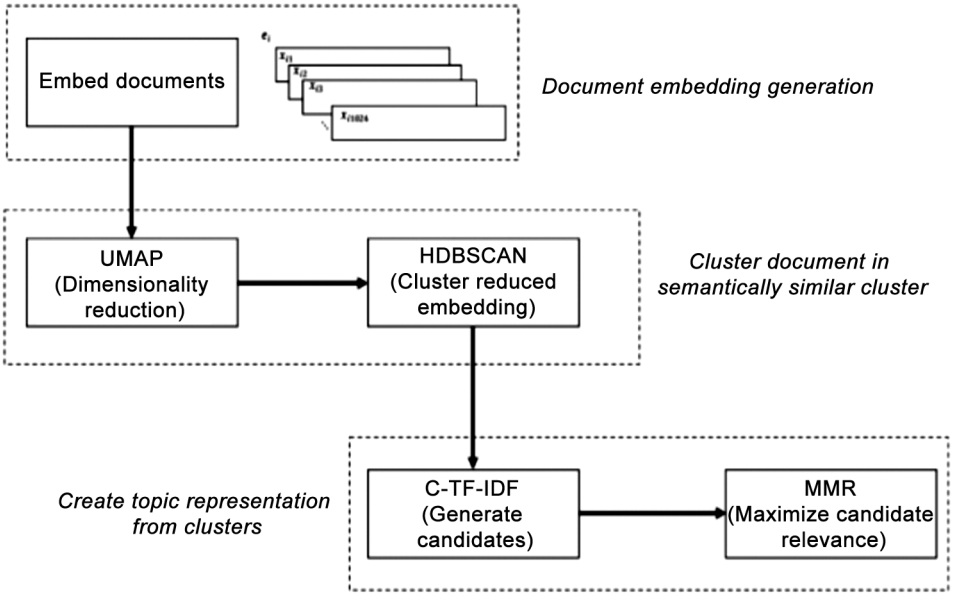

Fig. 2 — Clustering flow diagram

patents, emerging trends, and potential areas for future research. These findings were documented and presented in internal project reviews and external workshops.

Findings under Objective 3

To Support Continuous Tracking of Healthcare Innovations

Tool Development and Validation

To support the continuous tracking of healthcare innovations, a web-based application was developed. This tool integrates the machine learning models for real-time analysis of patent data. The application allows users to input new patent documents and receive insights on technological trends and classifications.

Tool Design

The design and development of the tool were undertaken by a team of six students, including both MSc and PhD candidates.

The process involved several stages:

- Requirements Gathering: Meetings with legal experts and potential users to gather requirements.
- Design and Development: Creation of a user-friendly interface and integration with the machine learning models.
- Testing: Rigorous testing to ensure the accuracy and usability of the tool. Policy Analysis:

The tool was used to analyse the impact of the National IPR Policy 2016 on healthcare innovations. This involved examining the changes in patent trends pre- and post-policy implementation. The analysis was performed by a research assistant under the supervision of Dr. Ashwini Siwal, who provided insights into the legal aspects of the policy.

Educational Deployment

The tool was also integrated into the university curriculum. Workshops were conducted to train students and faculty on its usage. Three workshops were held, each attended by around 30 participants. The sessions included hands-on training and discussions on the implications of patent analytics in healthcare.

Results

The tool is being planned for deployment for ongoing research and educational purposes. Initial feedback from users has been positive, highlighting the tool's effectiveness in providing valuable insights into patent data.

Discussion

The analysis of patent data using machine learning techniques revealed several new findings that significantly contribute to our understanding of technological trends in the healthcare sector. One of the key discoveries was the identification of emerging technology clusters that demonstrate significant innovation potential. These clusters were identified using a combination of BERT for text representation and LDA for topic modelling, followed by K-means clustering. The topic coherence scores, particularly with a minimum cluster size of 34, achieving a coherence score of 0.610, indicating strong semantic similarity within these clusters. The analysis highlighted the convergence of multiple technologies in public health, suggesting that future innovations are likely to stem from interdisciplinary approaches.

The project resulted in the development of a machine learning-based tool that can classify and cluster large volumes of patent data, providing real-time insights into technological trends. This tool, if integrated into a web-based application, allows users to input new patent documents and receive detailed analyses, facilitating continuous tracking of healthcare innovations. A combination of BERT and LDA (Annexure B) were previously not tested on textual patent datasets. The policy analysis component yielded valuable insights into the impact of the National IPR Policy 2016 on healthcare innovations, informing policy discussions and highlighting the need for supportive frameworks in public health technology adoption.

Conclusion

The project successfully developed a machine learning-based tool for patent analytics, significantly aiding public health innovation. The interdisciplinary collaboration between legal and technical experts yielded valuable insights and advanced the field of patent analytics in healthcare. The project culminates in deploying the machine learning framework within an academic setting, enhancing lectures, curriculum development, and policy-making processes. Consequently, our findings have implications in the domain of patents and healthcare categorically providing a way for enhancing the quality of patents in the long run. It is suggested that the challenges faced in conducting thorough examinations can be overcome by providing well-trained examiners with tools as developed in this research that enhance their ability to identify and address such issues.

References

1 Rai A K, Machine learning at the patent office: Lessons for patents and administrative law, *IOWA Law Review,* 5 (July) (2019) 2617.

2 Rai A K, SharmaI & Silcox C, Accountability, secrecy, and innovation in AI-enabled clinical decision software, *Journal of Law and the Biosciences*, 7 (1) (2020) 1, https://doi.org/10.1093/jlb/lsaa077.

3 Alcácer J, Gittelman M & Sampat B, Applicant and examiner citations in U.S. Patents: An overview and analysis, *Research Policy*, 38 (2) (2009) 415, https://doi.org/10.1016/j.respol. 2008.12.001.

4 Mann R J & Underweiser M, A new look at patent quality: Relating patent prosecution to validity, *Journal of Empirical Legal Studies*, 9 (1) (2012) 1.

5 Shalaby W & Wlodek Z, Patent retrieval: A literature review, *Knowledge and Information Systems*, 61 (2) (2019) 631, https://doi.org/10.1007/s10115-018-1322-7.

6 Vishnubhakat S & RaiA K, When biopharma meets software: Bioinformatics at the Patent Office, *Harvard Journal of Law and Technology,* 29 (1) (2015) 206.

7 Nicholson W & Rai A K, Clearing opacity through machine learning, *Iowa Law Review,* 106 (775) (2021) 775.

8 Govindarajan U H & Singh D K, forecasting cyber security threats landscape and associated technical trends in telehealth using bidirectional encoder representations from transformers (BERT), *Computers and Security*, 133 (Oct) (2023) 103404.

9 Govindarajan U H, Trappey A J &Trappey C V, Intelligent collaborative patent mining using excessive topic generation, *Advanced Engineering Informatics,* 42 (2019)100955.