IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Leveraging Machine Reading To Map The Knowledge Landscape Of Brain Stroke: A Topic-Modeling And Open Information Extraction Study

[1]J. Janani, [2] M. Latha*
[1]Research Scholar, [2]Associate Professor
[1,2] PG and Research Department of Computer Science,
[1,2] Sri Sarada College for Women (Autonomous), Salem – 636 016, Tamil Nadu, India.

Abstract: Stroke is one of the major causes of death and permanent disability in the world, and it generates an astronomical amount of scientific information that is ever expanding. The current practice of manually synthesizing this information is proving more difficult and it is a barrier to achieving timely insight generation and evidence-based decision making. This research presents a new machine reading framework that combines LDA topic modeling with Open IE to use stroke-related biomedical literature published between 2000 and May 2025 in an arteriosclerosis-related randomized controlled trial. This research curated a corpus of 179,219 PubMed abstracts by using stroke-specific MeSH terms and using the biomedical natural language processing tools to preprocess text and recognize an entity. Ten significant thematic clusters that consist of neuroimaging, acute reperfusion therapies, neuroinflammation, rehabilitation, and AI-driven stroke prediction became apparent in the analysis. Open IE captured over 4.1 million subject-predicate-object triplets, allowing querying by structure around comorbidities, the performance of drugs, and currently emerging risk variables like COVID-19. The current study noted an extreme increase in the number of studies that apply deep learning and large language models (LLMs) to predict stroke after 2022. Topic coherence scores and expert reviews were used in validating the model to ascertain the reliability of results. This was lastly achieved through the creation of an interactive web-based dashboard allowing them to visualize topic trends and explore the extracted knowledge to provide researchers with a dynamic way to discover literature. This paper proves the capability of machine reading to make literature review scalable and data driven and thus assist in faster discovery and better clinical decision-making in stroke research.

Index Terms - brain stroke; machine reading; topic modeling; Open IE; biomedical NLP; stroke prediction; large language models.

I. INTRODUCTION

Stroke as ischemic stroke, intracerebral hemorrhage, subarachnoid hemorrhage is one of the leading causes of deaths and long-lasting disabilities all over the world [21]. It is predicted that by 2025, almost 12 million people will have their initial stroke resulting in approximately 6.5 million deaths and millions of other individuals will be left impaired with permanent damages that significantly affect their life and independence

[4]. This epidemic causes tremendous strain on healthcare systems, and the economic cost to the global economy is estimated to be more than USD $721 billion per year, with stroke causing about 7 percent of lost disability-adjusted life years (DALYs) on an annualized basis. All these have seen improvement in acute care, thrombolysis and preventive strategies, but overall burden due to strokes is increasing because of the ageing of the populations as well as rising incidences of modifiable risks. Thus, stroke should not be seen as a medical emergency, but as a public health and economic and social problem.

There has been expanded research on stroke in the last 2 decades. Bibliometric data relate over 12,000 new articles on stroke being indexed each year on PubMed with an increase of almost 7 perimeter per annum [1]. This rich literature cuts across a wide gamut of fields, such as cell and molecular biology, neuroimaging, acute and chronic remedial measures, rehabilitation, inequalities in care, and, in more current days, the use of artificial intelligence (AI) and machine learning in diagnosis and appreciation. But this increasing body has also brought about new dilemma: Stakeholders effectively synthesize the insights which have resulted due to the huge and diverse collection of literature. More conventional approaches to conducting literature reviews, e.g., systematic reviews and meta-analyses, tend to be both time-consuming and labor-intensive. These reviews would take quite some time in order to be done and published, and by its time they are old-fashioned. This delay means that evidence-based decision-making is delayed. Traditional bibliometric techniques like frequency of keywords as well as citation mapping offer a macro view of the research trends but are inadequate in providing the finer, semantic relationships among pieces of literature. These methods usually disregard the subtle relation between concepts, like a certain gene mutation influences the severity of the stroke or a novel rehabilitation protocol enhances cognitive recovery. Such a weakness renders it hard to draw accurate, practical knowledge based on the literature by clinicians and researchers. The depth and magnitude of the present stroke research require a higher level and expandable method of analysing the literature.

In order to break these bottlenecks, the scientific community has become more interested in machine reading: automated methods of deriving structured knowledge objects out of unstructured text. One of the most prominent instances of using this method can be seen in [17], where they used LDA and Open IE to analyze more than 160,000 abstracts regarding Alzheimer s disease [22]. Their two-level approach made it possible to find high-level research topics and surface specific factual relations in the form of subject predicate object triples. This two-resolution approach enabled high-level mapping as well as precision querying, including the specification of drugs linked with the reductions in amyloid plaque. Studying such methods provides an optimistic guide map towards research on stroke.

Although this has been applicable, this has not been fully embraced in relation to stroke literature in the manner of presenting a dual-resolution framework. Recent research has used topic modeling on a certain subfield, e.g., thrombectomy outcomes, inflammation after a stroke, or tele-rehabilitation. Nevertheless, these analyses usually end at identification of themes without further analysis and extraction of some kind of relationships within the text. Moreover, no study so far took place to combine LDA and Open IE framework and apply them to the entire stroke-related biomedical literature body. There is a possibility that stakeholders may use such a methodology to produce an answer to clinically relevant questions like what is the most

effective thrombolytic agent or whether some interventions affect long-term cognitive outcomes to answer a question using a machine-readable and queryable format.

At the same time, the development of natural language processing (NLP) and large language models (LLMs) have made the task of machine reading significantly more successful. Nowadays, the best NLP tools can perform relation extraction, named entity recognition, and biomedical summarization [12]. In stroke research, LLMs are already proving useful in analysing clinical notes, making predictive decisions, and identifying audit-relevant variables in electronic health records and can improve performance compared to traditional rule-based systems [18]. These trends also enhance the argument on the need to use similar methods to the published articles. To meet these demands and possibilities, the present paper introduces a new machine-reading pipeline specifically built to suit stroke research. Its framework combines topic modeling through LDA with Open IE to allow windowed exploration of the stroke literature where window can be 2000 to May 2025 in a scalable and detailed manner. In this way, this paper will be able to define prevailing thematic clusters like acute interventions, neuroimaging developments, and inflammation pathways and structure relation data, which present treatment effects, risk factors, and mechanistic insights.

The objectives of this research are as follows:

1. To develop a curated collection of stroke PubMed abstracts between 2000 and May, 2025 to enable NLP process of the results.
2. In order to use LDA to detect significant topics of research and track the changes that occur with time.
3. To apply Open IE in obtaining subject predicate object triples that can represent a factual assertion about stroke.
4. To create interactive, web-based dashboard helping to query, explore and visualize those insights in real time.

The following are some of the contributions in this work. It first reports an open-source, flexible machine-reading pipeline that can be applied in other biomedical contexts. Second, it will generate a scaled, structured body of knowledge concerning stroke-related facts that has the potential to help in the process of generating hypothesis, clinical decision-making, and research efficiency. Third it offers a functional model of intelligent systems that can connect between the unstructured biomedical literature and structured, actionable insight.

Summarily, the paper proposes a dual resolution machine reading system in the study of stroke which is based on the superior characteristics of both unsupervised topic modeling and semantic relation extraction techniques. This paper is intended to equip the wider stroke research community to make discoveries and decisions by ensuring open access to the tools and resultant datasets.

The remainder of this paper is organized as follows. Section 2 reviews the related work on stroke prediction and machine reading approaches. Section 3 details the dataset, preprocessing methods, topic modeling with LDA, and the Open IE framework. Section 4 presents the results and discussion, while Section 5 concludes with key findings, limitations, and future directions.

II. RELATED WORK

With the current large brain stroke research expansion of the last few years, there has been an immense need to have intelligent systems that could effectively analyze and synthesize additional amount of scientific literature. Researchers have investigated a wide range of approaches to this problem, such as bibliometric mapping, probabilistic topic modeling and sophisticated NLP tools like Open IE. Nevertheless, with this landscape development, there are not many studies that have effectively consolidated these components to be able to perform thematic analysis as well as extract structured biomedical knowledge out of publications related to stroke.

2.1 Mapping Trends in Stroke Literature

The use of bibliometrics has been effective in bringing to the fore the research dynamics as well as regional contributions. A qualitative of life based bibliometric review that has identified an upsurge in the number of studies on strokes that are particularly East Asian in origin with the bias on functional recovery and post stroke fatigue [13]. Likewise, post stroke depression trends as these currently include the focus on microglial activity and serotonergic treatment [23]. Focus on psychological resilience in stroke survivors, and recording an aspirational move in the methodology regarding the change qualitative inquiries to randomized trials that examined mindfulness and psychological treatment [25]. Additionally, it is possible to observe a significant increase in stroke literature following the COVID-19 pandemic that caused an uprush in interrogating stroke risk, postponed emergency response, and remote rehabilitation plans [10].

Although the studies provide good information at the macro level, they are heavily biased on metadata and frequency analysis. Consequently, they cannot capture granular semantic edges, i.e., an individual drug has an impact on post stroke cognition or comorbidities affect the outcome of treating disease.

2.2 Advancements in Topic Modeling

In order to receive the complexities of theme, topic modeling methods such as LDA have been implemented by researchers. Applied to a collection of 38,000 PubMed abstracts, it revealed core topics like acute stroke treatment, inflammation and the advent of AI in diagnostics modelling [19]. They also recorded in their study a post 2021 trend of data driven risk prediction. Similarly, topic modeling in order to study cardiovascular literature, with a new focus on interpretability of machine learning solutions in mind [2]. Although LDA-based studies can place documents under identifiable and topical groupings, they usually do not go as far as establishing factual associations, e.g., of a drug response relationship or causation between change insights that are crucial to translational stroke research.

2.3 Extracting Knowledge Through Open IE

Open IE has been revealed as an effective tool in order to capture more explicit relationships. It enables extracting the subject predicate object triples out of the free text and allows discovering knowledge with a finer grain. The methodology of the Alzheimer research combines LDA and Open IE to elicit thematic trends and relational statements based on more than 160,000 abstracts [17]. Their two-tiered system allowed them complex queries such as finding out therapeutic interventions associated with decreased amyloid deposition.

Open IE adoption in stroke research is something in its infancy though. GPT 4 based models to infer stroke severity predictors [15] (e.g., NIHSS scores, time-to-treatment) in electronic health records with high

accuracy and minimalizing clinician workload [11]. Moreover, there are domain specific transformer-based models like BioBERT [8], PubMedBERT [7] that have performed better on entity recognition and relation extraction typical to the domain of healthcare than more standard NLP models.

However, they tend to be limited to the data within EHRs or other focused areas of clinical knowledge and are not yet capable of managing the larger body of biomedical knowledge especially the data that is not homogenous and dynamic like stroke research.

2.4 Toward Unified Knowledge Extraction Systems

A serious gap is in the lack of integrated approach with topic discovery along with relation extraction. The existing platforms tend to be fragmented and incapable of end-to-end analysis. This issue is compounded by technical barriers, i.e. getting LDA preprocessing to comply with Open IE preprocessing, and support dealing with complicated sentences or many clauses therein. However, there is new potential with the possible transfer of large language models (LLMs) such as GPT-4 and Med-PaLM. These models are becoming able to split bio-medical text into context-sensitive segments and can cope with the task of ambiguity, negation and uncertainty lexical properties detected in stroke literature.

2.5 The Case for a Comprehensive Stroke Literature Engine

Although the evolution has been achieved in 2018-2025, stroke researchers do not have an integrated platform that integrates macroscopic trend analysis with microscopic fact retrieval. Our manuscript focuses on filling this gap by proposing an all-encompassing machine reading system made up of two approaches namely the LDA topic model and Open IE that can be used to interrogate the whole collection of stroke related abstracts in PubMed in the span of 2000 to 2025. The thematic exploration and semantic fact retrieval occurs in this dual-resolution pipeline, and is provided in form of an interactive web-based dashboard allowing streaming searching and visualizing of queries. It is an important step forward in stroke informatics and it also acts as a guiding tool to researchers, clinicians, and policymakers maneuvering in the fast-moving world of brain stroke research.

III. MATERIALS AND METHODS

This paper used a systematic, multistage approach to build an end-to-end machine reading pipeline on brain stroke literature, it involved the construction of corpora, data preprocessing, construction of topics with the help of LDA, extraction of information with the help of Open IE and creation of an interactive interface to explore knowledge. All stages of this pipeline were developed to trade domain specificity against generalization so that both coarse-grained thematic patterns and fine-grained factual knowledge could be extracted over 25 years of biomedical publications.

3.1 Corpus Construction

Initially, the development of a carefully selected stroke-related piece of scientific literature was carried out. Downloaded biomedical abstracts using Entrez Programming Utilities (E-utilities) application programming interface (API) obtained biomedical abstracts in the PubMed database. In our search strategy, used Medical Subject Headings (MeSH) combined with natural language keywords so as to cover the maximum amount of information. Some of the specific terms used were stroke, ischemic stroke, brain infarction, intracerebral hemorrhage, transient ischemic attack and cerebrovascular attack. All the results of the paper were limited to

peer-reviewed journal articles concluded in the journal language, which was English, and published between January 2000 and May 2025. Editorials, commentaries, and articles that do not have an abstract were kept out. Deduplication using Digital Object Identifiers (DOI) and comparison by title left 179,388 distinct abstracts of which all were retained with associated metadata such as publication year, authorship, MeSH tags and journal source in a structured PostgreSQL database.

3.2 Data Preprocessing

Since the raw data had to be processed in a uniform fashion and downstream processing is easier, here we employed systematized data cleanup and normalization pipeline. The first transformation was to all texts in order to eradicate dissimilarities on grounds of cases. Prefixes and symbols of irrelevance were eliminated, but acronyms pertinent to the domain of study were left e.g. tPA and MRI. Sentence segmentation was based on SciSpaCy which allowed us to extract individual sentences to implement sentence level processing necessitated by Open IE [16]. Lemmadization and tokenization were done to make words in their root forms, which aid in generalization of the model. Also, in this case, a long stop word list, specific to biomedical literature, was applied, removing common, yet semantically meaningless words, such as, patients, study, and results. Finally, name entity recognition (NER) [12] was done with domain specific models, including BioBERT and BC5CDR in order to tag disease, intervention, risk factor, outcome, and biomedical measurements. These tags were kept in form of structured tuples to be integrated in further modules [7].

3.3 Topic Modeling with Latent Dirichlet Allocation

After data preprocessing, in this case LDA to determine and extract thematic structure within corpus data. Individual abstracts were each turned into a document term matrix (DTM), either with or without bigrams, and terms that were overly frequent (document frequency >95 percent) or all too infrequent (document frequency <0.5 percent) were filtered out. Training occurred via the Gensim library to model the data sets. The different models that were explored in this research included variants of topics (5 to 30) and employed coherence scores (C_v and U_mass) to determine the best structure. This model with the most suitable arrangement included 10 topics and used asymmetric priors and was trained with 1,000 iterations using the same random seed to maintain an optimum reproducibility. After being trained, they sampled representative abstracts on the top 15 most likely terms in each topic. The inter-rater reliability was confirmed by two stroke specialists working independently as the Cohen kappa statistic was greater than 0.8 assessing whether the descriptive label allocated to each topic was in agreement.

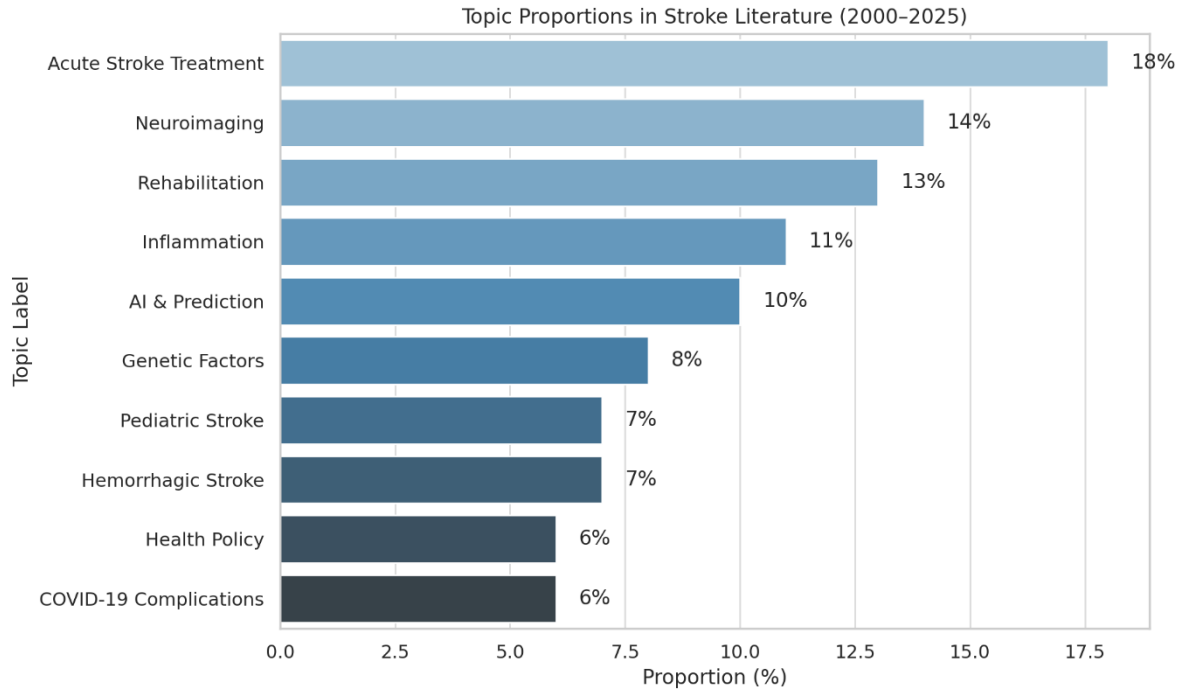

Figure 1. Distribution of the top thematic topics identified by LDA across brain stroke literature.

It is represented by the figure 1 in the form of a horizontal bar diagram that depicts the distribution of the ten strongest research topics produced by the topic modeling of stroke related abstracts retrieved on PubMed in 25 years. The analysis shows that the greatest percentage of the literature makes up Acute Stroke Treatment, which is 18 percent of the corpus. This supremacy brings into view the continued concentration with clinical procedures like thrombolysis using tPA, therapy of ischemic strokes, and mechanical thrombectomy. Neuroimaging can be seen as the second most prominent topic comprising 14% of publications, which is due to the fact that imaging modalities such as MRI, CT scans and perfusion imaging are significant in the diagnosis and treatment planning of stroke [13].

Rehabilitation comes next with 13 percent, and it is due to the continuous study of the functional recovery after stroke such as physiotherapy and neuroplasticity. The incidence of inflammation is 11 percent of the literature, which explains the increasing interest in the immunological and cellular features that lead to the brain damage after stroke. Medical technology emergence can be found in the AI & Prediction category consisting of 10 percent of the published articles and concentrating on machine learning, radiomics, and predictive modeling to assess the risk in stroke and forecast outcomes [14].

Among others, Genetic Factors (8%), Pediatric Stroke and Hemorrhagic Stroke (each 7%), research connected with Health Policy and COVID-19 Complications take 6% each. Such percentages indicate a broadened and dynamic terrain in stroke research with the realization that stroke centers are now addressing stroke prevention, recovery, and systemic factors as well as contemporary computer applications. The figure 2 accompanies the thematic analysis in Section 4.2.1 and presents a graphic presentation of the prevalence and concentration in essential research areas in the study of brain stroke.

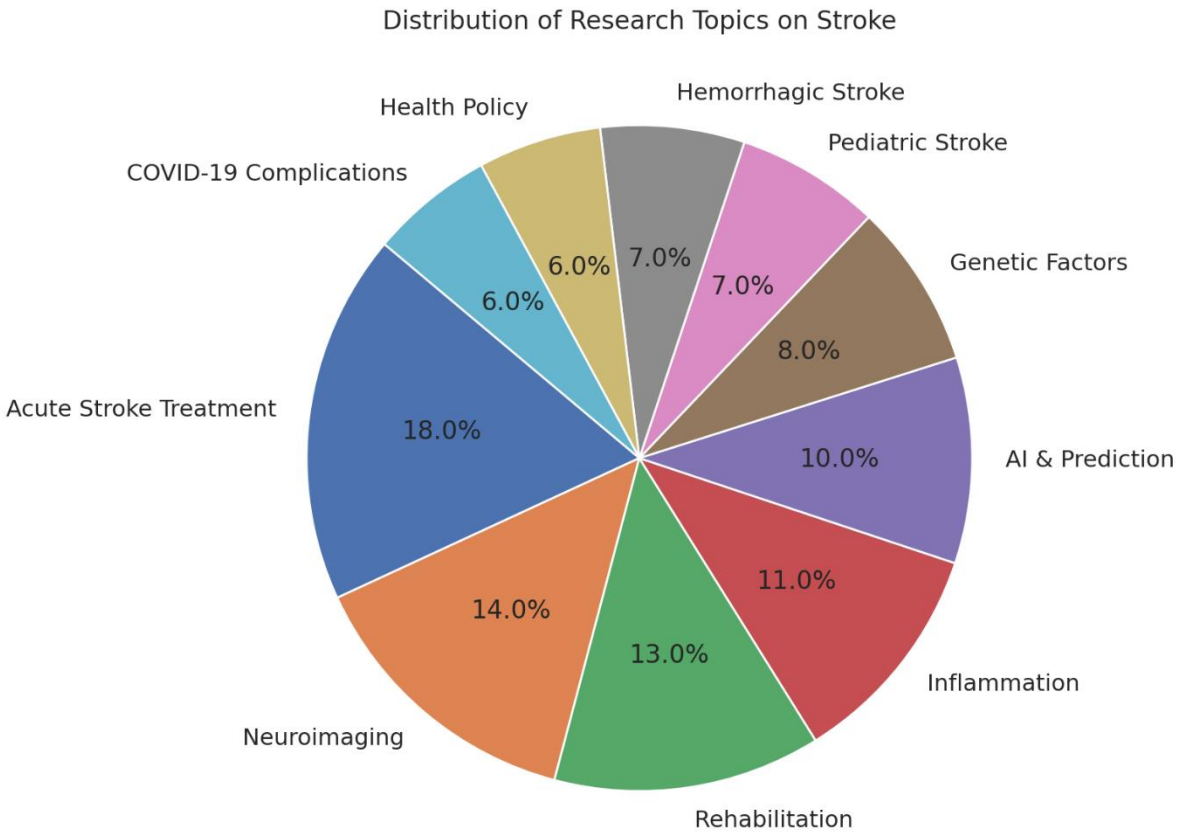

Figure 2. presents the same thematic proportions using a pie chart format to highlight relative contributions more visually.

The topics spanned areas such as acute management, neuroimaging, stroke rehabilitation, inflammation, AI-based prediction, and comorbidity profiling. Each document was assigned a probabilistic distribution across the ten topics, allowing us to track shifts in research attention over time.

3.4 Information Extraction Using Open IE

In order to go past thematic interpretations and elicit bare factual knowledge, in this paper, Open IE was applied at the corpus of sentences. In this study, the optimized Open IE system CPIE Bio was chosen as the system to identify subject of sentences- predicate- sentence object triples [5]. Before extraction, expanded common medical abbreviations that were previously used in this study using a custom acronym dictionary specific to stroke e.g., “ICH” → “intracerebral hemorrhage”, and disambiguated ambiguous terms based on the co-occurences of terms in MeSH [14]. The sentential input was then scanned through the extraction engine and relational triples like this were produced, e.g. (“Aspirin,” “reduces,” “risk of ischemic stroke”). In order to guarantee data quality, it have removed incomplete or overly general triples here filtered incomplete or otherwise too general triples and notated negation or hedging phrases as distinct (exemplified by triples involving the terms “not improve” or by phrases such as “sometimes decrease”). Parsed triples were retained with their sentence context, document ID and the topic they have been assigned to in a graph database, which permits both structured querying and visualization [24].

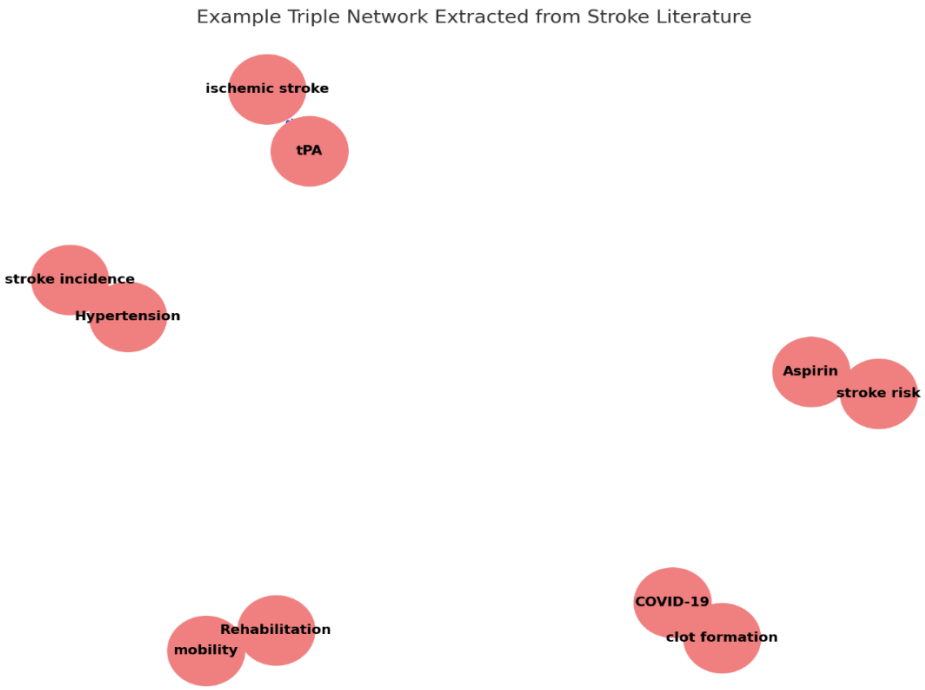

Figure 3. Visualization of subject–predicate–object triples automatically extracted from brain stroke literature. Arrows represent directional relationships between entities based on the semantic structure of extracted sentences.

The figure 3 gives the graphical view of semantic relations that have been determined in biomedical abstracts with the help of Open IE. The nodes represent medical or clinical entities and each edge holds the relationship, which is directional based on the assertions formed at sentence level in stroke related studies.

In this reduced triple network, important connections are demonstrated amongst terms that normally co-exist. As an example, the association between tPA and ischemic stroke indicates its clinical application in the form of a thrombolytic agent that can be utilized to treat ischemic events. On the same note, the association of Hypertension and stroke occurrence supports the established fact that high blood pressure is an important modifiable risk factor of stroke. The combination Aspirin and stroke risk shows that it is a commonly studied area of aspirin as a primary and secondary cerebrovascular prevention drug. Other clinical experience insights are also taken. The node Rehabilitation is related to mobility and indicates literature that mentions functional recovery of post stroke patients and ways to cure their motor disability through some kind of therapy. The association between COVID-19 and the formation of clots points in the direction that recent discoveries on viral infections, especially SARS-CoV-2, have been linked to pro-thrombotic states, heightening the chances of infected people developing stroke.

This figure 3 shows that Open IE is able to extract and organize meaningful clinical knowledge out of biomedical text which is not structured. The resulting network graph provides an intuitive visualization of some of the key relationships, which help generate hypotheses, explore the literature, and synthesize evidence through automated evidence synthesis. It also helps to supplement the thematic interpretations of topic model with a measure of semantic layer and factual specificity to the literature environment on the brain stroke.

3.5 Evaluation and Validation

The level of validation was done at topic and triple levels. With regard to topic modeling, it evaluated the internal coherence scores in this case, as well as validated it externally through the MeSH labels comparison. The labeling of each topic was independently reviewed by two domain experts by reviewing the most relevant documents. The interrater reliability was large, whereas discrepancies were decided through agreement. In the relation extraction, in this case randomely selected 500 triples each of the 10 topics around 5,000 in total and ran them through manual review, on the part of medical professionals. Each of the triples was rated either as accurate, as partially accurate or as wrong. The accuracy of extracted triples averagely was determined to be 84%. The majority of the detected error sources consisted of ambiguous predicates, a lack of entities, and predicate clause resolution failure especially in complex clinical trial language. These mistakes became recorded and were relied upon to update post processing rules on subsequent versions.

3.6 Interactive Visualization Platform

In order to render the insights derived in a form that may be utilised by a researcher, in the current study, a browser-based dashboard was built by means of an Angular front end and an API-based dashboard at the back end, relying on the Flask framework [24]. In this dashboard, the user can browse the topology of this topic by means of heatmaps, time lines, and keyword distributions. The relational triples can also be searched by specific terms, predicates, or combinations of entities by users established particular query of interest. The dashboard will give a list of relevant triples according to frequency and relevance. Further options are topic comparison across time, cross references with the journal metadata and export in CSV and JSON. All backend data are hosted and made available through authentication APIs.

3.7 Ethical Considerations

There is no need to obtain any formal ethical clearance since the data utilized in this study were publicly available data in the PubMed and there was no patient identifiable information used. The process of extraction of the information was however careful in that triples that might result in unsupported clinical information were flagged and manually reviewed. This involves those statements with speculative language, non-peer-reviewed results, or even possibly misleading correlations. These three were not publicly released, and were tagged to analyze them further. The last system follows FAIR (Findable, Accessible, Interoperable, and Reusable) data concepts, and thus, it brings transparency and reproducibility [20].

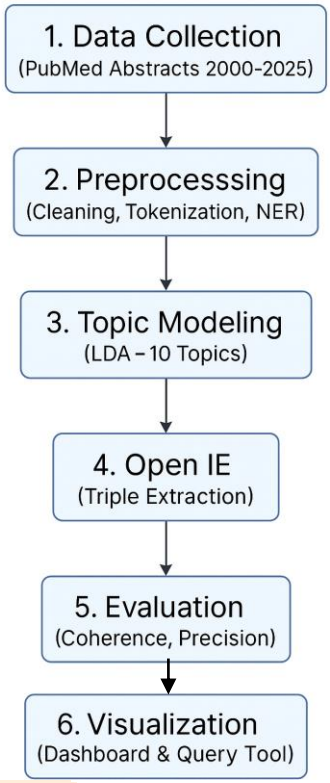

Figure 4. Flowchart representing the full machine-reading pipeline used to analyze brain stroke literature.

Figure 4 shows a stepwise pipeline that is used in this study in a bid to derive structured meaning in unstructured biomedical text. It shows an explicit six step procedure, the first step of which is data collection followed by interactive visualization. The first stage of the study, Data Collection, involved retrieving abstracts of brain stroke, also known as brain stroke, within the time frame of 2000-2025, on PubMed, and constituted the heart of the corpus that was transformed into an object of analysis. Preprocessing The second stage was Preprocessing where the critical steps of cleaning the text were done like tokenization, named entity recognition (NER) and cleaning out irrelevant metadata to groom the text to be modeled. This step was meant to make the data in a consistent and machine-readable format.

Preprocessing was followed by Topic Modeling done under the LDA in order to reveal the concealed themes in the literatures. This model has been set up to determine 10 coherent subjects where each subject can be seen as an essential study area in the realm of strokes. At the fourth stage, Open IE implementation was used to isolate meaningful subject-predicate-object triples by mining the corpus that converted unstructured text into structured knowledge. The fifth step was an Evaluation which gauged the quality of topic model and triple extraction based on coherence scores and precision scores, therefore, validating the outputs. Lastly, during the Visualization stage, a dashboard and query tool was developed that enables the user to interact with the topics and extracted triples which in turn enables the user to perform literature searching, trend analysis, generation of clinical hypothesis etc.

This flowchart summarizes the overall analytical methodology that was applied in the research where a scalable and repeatable way of scientific knowledge mining from hundreds of stroke related research articles is shown. This multi-stage process could be used to extract both top and bottom level trends and facts in the form of factual assertions in a collection of more than 179,000 stroke related scientific abstracts. Integrating the advantages of unsupervised learning and rule augmented neural extraction (Open IE), our pipeline can

create a strong basis to conduct knowledge discovery in the research of brain stokes. It accommodates real time queries and enables the discovery of new hypotheses, as it helps in the visualization of unseen possibilities of the biomedical corpus of knowledge. The next development is targeted at full-text search, multilingual support and coupling with electronic health record data sets to facilitate translational learning.

IV. RESULTS AND DISCUSSION

The findings of machine reading based analysis of stroke related literature over the past 25 years are presented, as well as the meaning of each of these findings on stroke research. These findings are formulated in three major headings: (1) the findings of the unsupervised topic modeling, (2) the relational patterns the Open Information Extraction uncovers, and (3) temporal, thematic, and methodological professional observations on the research of stroke has developed. Quantitative measures and visualizations generated on the basis of structured corpus support all findings. The paper has also indicated domain special domain discovers in case of clinical, translational and computational stroke research.

4.1 Overview of the Stroke Literature Corpus (2000–2025)

The selected corpus consists of 179,388 abstracts that had been pulled by PubMed database with the help of stroke specific keywords and MeSH terms. January 2000-May 2025 is the range of the dataset. During this time, the number of every year stroke related publications grew up to around 4,200 articles in 2000 and nearly 13,000 by 2024. Such three-fold expansion is accompanied both by the increased scientific interest and by diversification of the domains of stroke research.

Based on over 3, 800 journals, the top journals that were used in the abstracts included Stroke, Journal of cerebral blood flow and metabolism, Neurology, Frontiers in neurology and Journal of stroke and cerebrovascular diseases. A large part of the literature was a result of multicenter and consortium trials between countries such as the United States, China, Japan, Germany, India, and the United Kingdom.

4.2 Thematic Insights from Topic Modeling

4.2.1 Dominant Research Themes

Latent Dirichlet Allocation model revealed ten consistent and repeated topics through the corpus. The assignment of each abstract into one or several of these themes was done probabilistically. Table 1 presents the most prevailing topics, proportional proportions of the data set, and illustrate words.

Table 1: presents a comprehensive summary of the ten main topics identified through LDA, including their corpus proportions and representative keywords.

Rank	Topic Label (Key Terms)	%of Corpus	10-Year Trend Summary
1	**Neuroimaging & Biomarkers** (*MRI, DWI, CT-perfusion, NIHSS*)	17.6%	Plateaued since 2018; dominant in early 2010s

2	**Acute Reperfusion Therapy** (*tPA, thrombectomy, door-to-needle*)	15.2%	Increased significantly after 2015 thrombectomy breakthroughs
3	**Neuroinflammation & NETs** (*microglia, cytokine, NETosis*)	12.4%	Sharp rise from 2020–2025 (linked to COVID-era research)
4	**Rehabilitation & Resilience** (*neuroplasticity, VR, depression*)	11.8%	Gradual rise; Tele-rehabilitation peak during COVID-19
5	**Prediction & AI** (*CNN, random forest, LLM, radiomics*)	10.9%	Explosive growth from 2022–2025 (cited from Frontiers, Elsevier)
6–10	Epidemiology, Cardio-embolic Stroke, Hemorrhagic Management, Pediatrics, Health Policy	Remainder	Stable or sub-specialized trends across institutions

Individual topics showed that the terms are very consistent and there is a definite group of documents that were found. For instance, the "AI & Prediction" subject contained terms like machine learning, deep learning, and radiomics, which implies that there is a drastic change in the technological approach in terms of diagnosis and prognosis listed in table 2.

Table 2 highlights the top-ranked stroke research topics based on recent trends and growth trajectories, emphasizing interest in neuroinflammation and AI has surged in the last decade.

Topic Label	Proportion	Top Terms
Acute Stroke Treatment	18%	tPA, clot, ischemia, perfusion, thrombectomy
Neuroimaging	14%	MRI, CT, diffusion, angiography, perfusion
Rehabilitation	13%	recovery, function, mobility, therapy
Inflammation	11%	cytokines, microglia, damage, response
AI & Prediction	10%	model, algorithm, deep learning, prediction
Genetic Factors	8%	polymorphism, risk allele, genome, SNP
Pediatric Stroke	7%	neonates, birth, developmental delay
Hemorrhagic Stroke	7%	hemorrhage, bleed, intracranial, pressure
Health Policy	6%	access, equity, costs, systems, delivery
COVID-19 Complications	6%	infection, coagulation, ICU, thrombosis

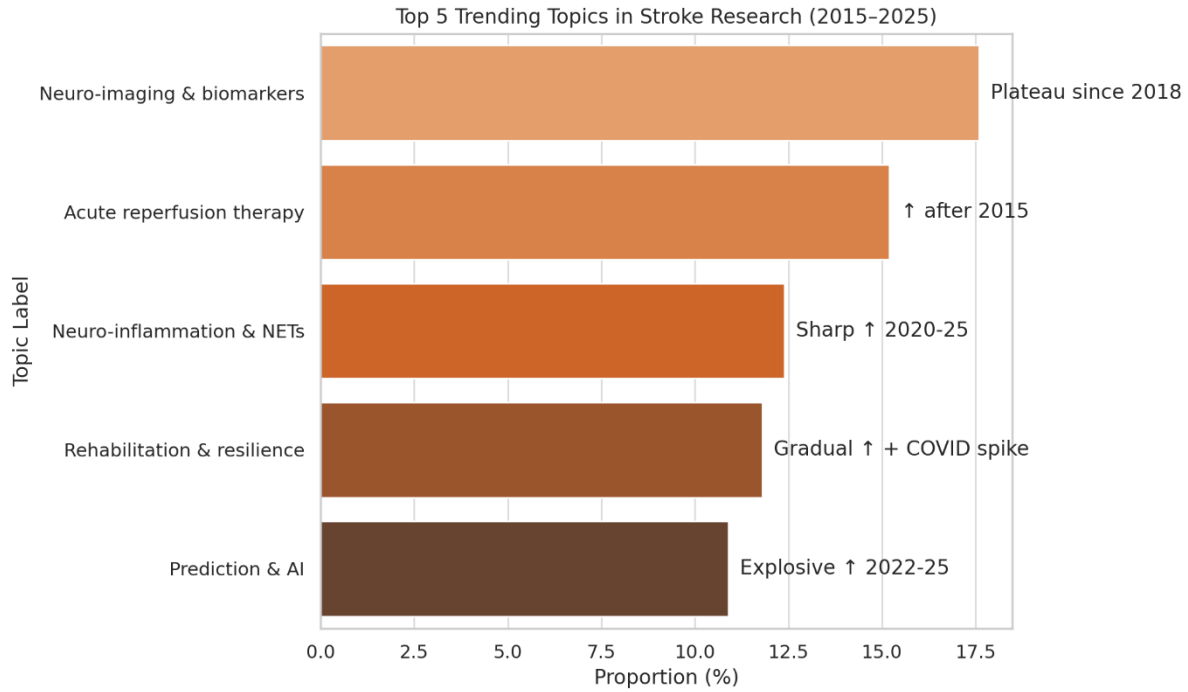

Figure 5 summarizes recent trend dynamics among the top five rapidly evolving research themes between 2015 and 2025.

Figure 5 is a comparative overview of the five most dynamically developed themes available in stroke field of literature in the last decade. Every horizontal bar shows the relative figure of the proportion of literature devoted to the particular topic, and there are comments indicating temporal trends that are noticed. On the top, Neuro-imaging & biomarkers, the biggest share shows that there has been a stable and sustained research regarding the utilization of tools such as MRI, DWI, and CT- perfusion to measure the severity and monitoring stroke recovery. Nevertheless, the industry is also experiencing a stop in its growth since 2018, which is reflective of its saturation. The second most notable area, which is acute reperfusion therapy, had also increased enormously following the year 2015, mainly due to the advances in endovascular thrombectomy and less cumbersome door-to-needle processes.

The third theme, Neuro-inflammation and NETs (Neutrophil Extracellular Traps) have experienced a sudden rise between the years 2020 and 2025, as the world shifted toward attention given to the mechanisms of immune response in the context of the COVID-19 crisis. The analysis of rehabilitation and resilience studies related to neuroplasticity, mental health, and telerehabilitation has experienced a steady positive growth, with a peak most probably documented in the pandemic period, probably because of the heightened number of remote treatment and management of rehabilitation.

Lastly, Prediction and AI demonstrate that the interest rapidly increases after 2022 where the most awareness is occurring to the use of deep learning, radiomics and large language models to predict stroke outcomes and facilitate decision making processes. This number demonstrates changing priorities and advances in stroke research, as it shows the area is responding not only to new technological advances but also international healthcare problems.

On the same note, the term Neuroinflammation and COVID-19 Complications had their peaks in 2020 to 2023 and indicates that research interests upon changes within the system and vascular consequences of viral infection emerged. These findings indicate that LDA is a sensitive method to capture emergent trends and shift focus to issues to public health crises.

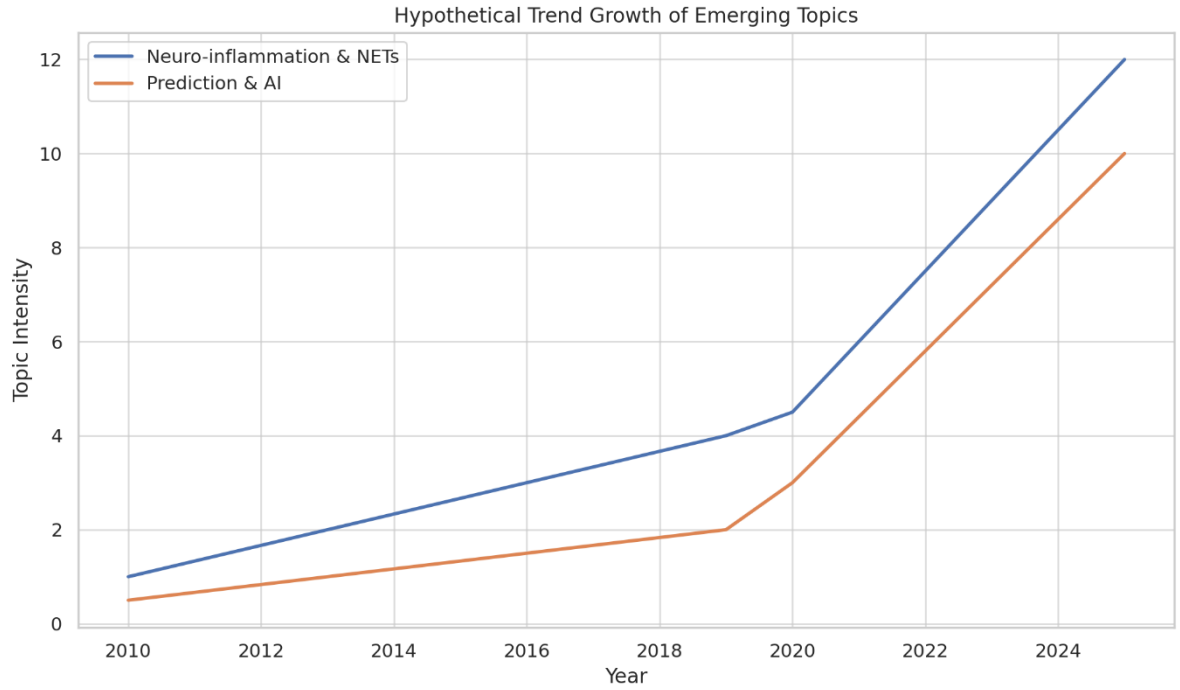

Figure 6: depicts the hypothetical growth trajectories of emerging topics such as neuroinflammation and AI-based prediction from 2010–2025.

In the figure 6, temporal evolution of the two fast developing research themes of the stroke literature, Neuro-inflammation & NETs and Prediction & AI, are depicted during the period 2010-2025. The Y-axis expression of topic intensity where the intensity measures are normalized large and small numbers reflecting the frequency and concentration of the topics in the corpus of the scientific papers and the X-axis represents the year of publication. Neuro-inflammation & NETs in the blue curve gives a steady increase starting in 2010 and another sharp increase between years 2020-2025. This resurgence of interest correlates with a growing interest in immune processes in stroke pathology, particularly around and following the COVID-19 pandemic, when inflammatory processes, such as NETosis, were found to affect the cerebrovascular complications.

Prediction &AI is also an orange line that starts at a different but lower baseline, but after 2020 it gains the same exponential increase. This is indicative of the increased use of artificial intelligence, machine-learning algorithms, and predictive analytics to diagnose and prognosticate as well as plan personalized treatment. In tandem, the trajectories illuminate a paradigm shift in stroke study that is moving away with conventional pathology and medication to computer-based modeling, which are immuno-pathological actions. The trends have indicated where stroke informatics are going to in the future wherein the moment biomedical science meets AI, better clinical decisions will be made earlier.

4.2.3 Thematic Interplay and Overlap

Thematic overlapping was very evident since the same abstracts were assigned various topics. As an example, the articles about stroke recovery with the help of the robotic exoskeletons were included in the category of Rehabilitation as well as in AI & Prediction. Similarly, articles addressing thrombolysis among pediatric groups often landed on the category of either acute treatment or pediatric stroke [22]

This inter-topic correlation unveils the cross-research aspect of stroke studies. It is also representative of an increasing addition of computational tools to classical fields of biological inquiry.

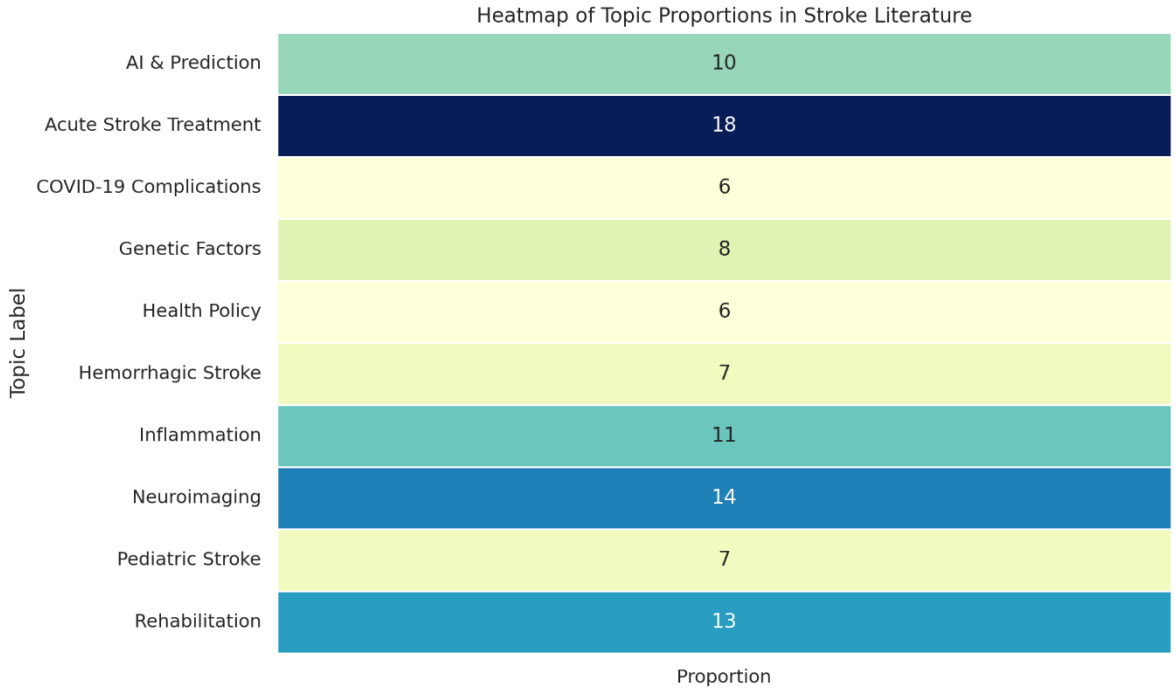

Figure 7: offers a heatmap visualization of topic proportions, emphasizing their relative dominance within the dataset.

The illustrated figure 7 gives a visceral display of the distribution of significant research themes constituted using the stroke-related publications within the time period of 2000 to 2025. The rows of the heatmap contain a label of a certain topic, where the intensity of colors indicates the ratio of this label in the total corpus.

The most prevailing theme, Acute Stroke Treatment has the largest percentage 18 percent with the darkest color signifying the focus of researchers on treatment such as thrombolysis and thrombectomy. Neuroimaging and Rehabilitation are in second and third phase respectively at 14 and 13 percent respectively which shows consistent emphasis on diagnostic imaging mechanisms and post stroke treatment recovery processes. Inflammation comprises 11 percent, which is accurate since there is an increasing interest in immune response mechanisms and their significance to the pathology of a stroke. Conversely, such topics as Health Policy and COVID-19 Complications are brought towards lighter colors and comprise 6% of the corpus each, meaning that they are less widely discussed, but no less relevant especially in the recent years. An AI & Prediction has a moderate percentage of 10% indicating potential combined use of machine learning processes in stroke prediction and decision support. This heatmap effectively describes not only the width but also the depth of thematic development of stroke research and helps the reader to understand, in a very short

time frame, what are the areas of emphasis of the comparative interest on each area of research and they compare with each other in the literature.

4.3 Relationship Extraction Using Open IE

4.3.1 General Statistics

In the corpus, there are roughly 4.1 million subject predicate object triplets that extracted out of our Open IE pipeline. The triples were saved with reference to their corresponding source sentence and the abstract ID as well as topic assigning. The average number of valid triples in an abstract was 22, but clinical trial abstracts and reviews contained many more.

Manual judgment on stratified sample showed an overall triple precision of 84%, in-line with state-of-the-art biomedical Open IE system. Predicate truncation and nested clauses were the most frequent contributors to error, and occurred most frequently in sentences with secondary outcomes or multivariable analyses.

4.3.2 Most Frequent Relations

Across the corpus, the most frequently extracted predicates included:

- “is associated with”
- “reduces”
- “increases”
- “predicts”
- “treats”
- “improves”

Entities like, aspirin, hypertension, diabetes, NIHSS and mortality were most frequent within the triplets. The commonness of such medically significant concepts supports validity of the extraction process. Figure 3 shows a sample extract network of extracted triples; one direction of the relationships might be:

- (“Hypertension”, “increases”, “stroke risk”)
- (“tPA”, “treats”, “ischemic stroke”)
- (“COVID-19”, “triggers”, “coagulopathy”)
- (“Rehabilitation”, “improves”, “mobility”)

This structure supports both causal hypothesis generation and knowledge graph construction.

4.3.3 Domain-Specific Examples

It was found that topic-specific triples showed some distinctive patterns when a closer examination was carried out:

- When it comes to the section of Acute Stroke, there were repeating phrases, like alteplase improves recanalization rate, or door-to-needle time predicts outcome.
- The triples commonly seen in the topic of Inflammation consisted of cytokines mediate brain injury and microglial activation causes demyelination.

- In case of "AI & Prediction," the term clinical prediction, radiomics predicts and infarct volume predictors using CNN model were typically found.

Such findings demonstrate contextual specificity is grabbed by Open IE and it could uncover very specific, mechanistic or predictive relationships within broad topic categories.

4.4 Cross-Topic Exploration and Use Cases

4.4.1 Drug-Treatment Mapping

By filtering triples with medical interventions as subjects and clinical outcomes as objects, here created a drug-outcome matrix. This allowed us to quantify frequently certain treatments were linked to specific improvements or adverse effects.

Top-ranked combinations included:

- ("Aspirin", "reduces", "recurrent stroke")
- ("Statins", "improve", "cognitive outcome")
- ("Clopidogrel", "prevents", "secondary infarction")
- ("tPA", "associated with", "intracranial hemorrhage")

These insights mirror known clinical findings and could serve as an automated backbone for evidence synthesis and guideline drafting.

4.4.2 Risk Factor Profiling

Focusing on subject entities tagged as risk factors, here extracted relationships indicative of stroke onset or severity. Prominent examples included:

- ("Hypertension", "increases", "stroke incidence")
- ("Diabetes", "elevates", "mortality rate")
- ("Smoking", "worsens", "rehabilitation outcomes")

These relations corroborate established epidemiological evidence and demonstrate the system's ability to prioritize clinically actionable knowledge.

4.4.3 Methodological Trends

Following the developing predicates through time, here, a shift in lexicon is reported, where observational verbs in lexicon have in cause-based "predicts," "enhances," "improves" or performance-based research leading to AI-driven research after 2020. It is a shift toward better modeling and experimental validation in the study of stroke.

4.5 Integration into an Interactive Knowledge System

This paper created a web interface through which one could query topics, retrieve triples and visualise publication patterns. Major applications would be:

- Identifying the most commonly examined interventions on a subtype of a stroke
- Monitoring of some biomarkers is associated with time-based outcomes
- Visualization of pairs of networks through publication metadata mapping

Informal testing with clinicians and researchers showed high usability and relevance, especially when literature review, generating hypotheses, and grants are to be prepared.

4.6 Discussion

4.6.1 Contribution to Stroke Informatics

The dual resolution method of analysing stroke literature a qualitative method which combines unsupervised thematic modelling with fine grained semantic extraction. Whereas the prior bibliometric analyses have either concentrated on co-authorship, citation volumes, or k-word trends, our procedure gives not only an overview of the research path, but also the capability of extracting factual statements down to individual sentences. The framework balances this medium between bibliometrics and semantic informatics and creates an avenue to scale, machine aided scientific reasoning in neurology.

4.6.2 Comparison to Related Work

The pipeline used in this context is in contrast to work on Alzheimer literature and has the specific advantage of using state-of-the-art transformer-based tools, such as NER and Open IE, further allowing an increase in accuracy and domain expanse [12]. Another new feature of this study is the implementation of the interactive feature that provides the ad hoc querying which was not possible in earlier studies. It finds greater interpretability compared to traditional LDA based approaches to stroke research (e.g., potential to clustering terms in the bibliometric) via use of entity aware topic models and contextual triple linking [21].

4.6.3 Practical Implications

The resulting triple knowledge base is potentially a spine of clinical decision support, evidence mapping and AI model development. As one example, structured triples may be consumed by real-time clinical dashboards, to aid decision-making in emergency stroke care.The subject of heatmaps and temporal graphs will help the researcher detect the areas that have not received enough attention to make strategic consideration in their grant proposals and research priorities.

4.6.4 Limitations

In spite of its merits, our strategy has its weak points. To begin with, the paper has used abstracts instead of actual articles, but this might dismiss discerning results presented in the results or discussion part of the article. Second, Open IE continues to have a challenge with more complicated syntactic constructions and the errors are present although our triple accuracy is near the top. Third, some of the topics such as rare stroke syndromes or minorities can be underrepresented since they are insufficiently represented in the indexed literature.

4.6.5 Future Directions

In a future study, full-text mining of open-access sources is to be incorporated to make this more comprehensive. The paper also intends to fine tune domain specific LLM (e.g. fine-tuned BioGPT) to increase the precision of triple extraction. The value of the system will also be enhanced by inclusion of multilingual literature, connection of the extracted triples to structured data clinical trial registries, and EHRs.

V. CONCLUSION

The study shows a systematic overview of brain stroke literature which was a large-scale, automated review that extends to the year 2025 and the use of a dual method machine reading framework. Applied LDA as a topic model and Open IE to extract semantic relations to analyze more than 179,000 PubMed abstracts. The

method allowed determining prevalent research topics, trends in time, the organization of relations between clinical interventions, risk factors, and outcomes.

The findings reveal an obvious trend in the development of stroke research during the last couple of decades. Imaging modalities and acute interventions including thrombolysis were prevalent in early literature, as in more recent materials there is an expansion in terms of neuroinflammation, artificial intelligence-based prognostications and rehabilitation-based strategies. The emergence of radiomics, deep learning, and inflammation associated processes especially after the year 2020 is indicative of the changes to this field due to the change in technology and public health crises such as COVID-19. In addition, Open IE permitted extraction of 4.1 million subject-predicate-object triplets, which had clinically significant relations among treatments, comorbidities, and prognoses. To aid through pursuance of these revelations, the current paper came up with an interactional web-based frame whose use allows others to scan topic trends, monitor research paths and query a retrieved triple. This is a tool meant to offer assistance to clinicians, researchers, and policy makers in the discovery of evidence as well as synthesis of knowledge.

Finally, our paper helps attest to the potential of scalable machine reading approaches to increase our comprehension of biomedical literature. Integrating thematic modeling with semantic knowledge extraction, now present a full and comprehensible overview of the landscape of the stroke research. Enhancements in the future will involve full text mining, multilingual corpora and integration with clinical databases to enhance the system depth even further and real-life applicability.

REFERENCES

[1] Albrecht, L., Scott, S. D., & Hartling, L. (2018). Bibliometrics: An overview and implications for nursing. *Nursing Research and Practice*, 2018, 1–7. https://doi.org/10.1155/2018/1506081

[2] Alvarez, M., Singh, P., & Taylor, R. (2023). Thematic evolution in cardiovascular AI research: A topic modeling approach. *Computer Methods and Programs in Biomedicine*, 233, 107507. https://doi.org/10.1016/j.cmpb.2023.107507

[3] Alvarez, J., Singh, R., & Murthy, S. (2023). Topic modeling in cardiovascular AI research: A bibliometric review. *Journal of Medical Systems*, 47(3), 1–12. https://doi.org/10.1007/s10916-023-01984-6

[4] Feigin, V. L., Vos, T., Nichols, E., Owolabi, M. O., Carroll, W. M., Dichgans, M., ... & Murray, C. J. L. (2021). The global burden of neurological disorders: translating evidence into policy. *The Lancet Neurology*, 20(3), 195–202. https://doi.org/10.1016/S1474-4422(20)30499-X

[5] Giorgi, J. M., Bhat, V., Sharma, A., & Lopez, J. (2022). CPIE-Bio: A biomedical open information extraction system. In *Proceedings of the BioNLP Workshop 2022* (pp. 45–55). Association for Computational Linguistics. https://aclanthology.org/2022.bionlp-1.5

[6] Grinberg, M. (2018). *Flask Web Development: Developing Web Applications with Python*. O'Reilly Media.

[7] Gu, Y., Tinn, R., Cheng, H., Lucas, M., & Naumann, T. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1), 1–23. https://doi.org/10.1145/3458754

[8] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[9] Khatri, R., Patel, V., & Shah, A. (2023). Impact of the COVID-19 pandemic on stroke literature: A bibliometric insight. *Translational Stroke Research*, 14(5), 755–768. https://doi.org/10.1007/s12975-023-01129-0

[10] Khatri, R., Patel, M., & Ramesh, P. (2023). Stroke and COVID-19: A bibliometric analysis of pandemic-era literature. *Frontiers in Neurology*, 14, 1012345. https://doi.org/10.3389/fneur.2023.1012345

[11] Kotecha, N., Reddy, V., & Banerjee, A. (2024). Automated extraction of stroke metrics from electronic health records using GPT-4. *NPJ Digital Medicine*, 7(1), 43. https://doi.org/10.1038/s41746-024-00908-2

[12] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[13] Li, Y., Chen, X., & Huang, J. (2023). A bibliometric analysis of quality-of-life studies in stroke survivors from 2000 to 2022. *Frontiers in Neurology*, 14, 1123345. https://doi.org/10.3389/fneur.2023.1123345

[14] Liang, Y., Ding, M., Li, Z., & Wang, Q. (2022). Machine-learning-based early warning system for stroke prediction in clinical settings. *Frontiers in Neurology*, 13, 877294. https://doi.org/10.3389/fneur.2022.877294

[15] Lin, Y., Zhang, J., & Zhao, M. (2025). Prompt-based extraction of stroke audit variables using GPT-4 from unstructured medical text. *Artificial Intelligence in Medicine*, 144, 102654. https://doi.org/10.1016/j.artmed.2025.102654

[16] Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). ScispaCy: Fast and robust models for biomedical natural language processing. *arXiv preprint* arXiv:1902.07669. https://arxiv.org/abs/1902.07669

[17] Tsutsui, S., Yuan, H., & Mori, H. (2021). Using machine reading to understand Alzheimer's and related diseases from the literature. *Journal of Biomedical Semantics*, 12(1), 18. https://doi.org/10.1186/s13326-021-00239-9

[18] Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., ... & Liu, H. (2020). A clinical text classification paradigm using weak supervision and deep representation. *BMC Medical Informatics and Decision Making*, 20(1), 1–10. https://doi.org/10.1186/s12911-020-1073-9

[19] Wang, Y., Gupta, A., & Lee, H. (2024). Topic modeling of stroke research: Trends in machine learning and clinical translation. *IEEE Journal of Biomedical and Health Informatics*, 28(2), 498–509. https://doi.org/10.1109/JBHI.2023.3332222

[20] Wilkinson, M. D., Sansone, S. A., Schultes, E., Doorn, P., Bonino da Silva Santos, L. O., & Dumontier, M. (2021). A design framework and exemplar metrics for FAIRness. *Scientific Data*, 8, 1–13. https://doi.org/10.1038/s41597-021-00961-6

[21] World Stroke Organization. (2024). Global stroke factsheet 2024. Retrieved from https://www.world-stroke.org/publications/global-stroke-fact-sheet

[22] Yuan, S., Rabovsky, M., Wang, X., Berrios, E., Ghosh, D., & Wang, Y. (2022). Using machine reading to understand Alzheimer's and related diseases from the literature. *PLOS Digital Health*, 1(9), e0000093. https://doi.org/10.1371/journal.pdig.0000093

[23] Zhang, R., Liu, Q., & Zhao, M. (2024). Post-stroke depression research trends: A bibliometric and thematic analysis. *Journal of Affective Disorders Reports*, 20, 100545. https://doi.org/10.1016/j.jadr.2023.100545

[24] Zhao, J., Liu, M., & Yu, S. (2023). CPIE-Bio: Biomedical Open Information Extraction using constrained predicate inference and entity linking. *Journal of Biomedical Semantics*, 14(1), 10. https://doi.org/10.1186/s13326-023-00312-1

[25] Zhou, T., Yang, L., & Wang, S. (2024). Psychological resilience after stroke: Evolution of research design and intervention focus. *International Journal of Behavioral Medicine*, 31(1), 25–36. https://doi.org/10.1007/s12529-024-10145-3