International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online)) Volume 14, Issue 1, January 2025)

Machine Learning and NLP-Based Analysis of Ayurvedic Medicinal Systems

Shubham Kanungo[1], Shambhavi Jha[2]

[1]*Assistant Professors, Department of Mechanical Engineering, IPS Academy, Institute of Engineering and Science, Indore*

[2]*Scholar, Department of Computer Science Engineering (Data Science), IPS Academy, Institute of Engineering and Science, Indore*

***Abstract*— Across the last few decades, digitalisation has reconfigured the way ancient knowledge systems are interpreted and developed. Data Science, fueled by AI, is now a primary driver in transforming raw data into meaningful insights in numerous disciplines. In medicine, it has been particularly useful for rapid and accurate diagnosis and treatment. Ayurveda, an ancient Indian system of medicine renowned for its holistic and side-effect-free approach, is gaining increasing attention. This work delves into the ways Data Science can optimize Ayurvedic wisdom with machine learning and text mining. The research examines a vast database of Ayurvedic medicines to identify patterns, categorise similar drugs, and present findings. By clustering and employing TF-IDF methodology, the paper presents a new, data-oriented approach to categorising ancient Ayurvedic drugs—laying the foundation for extensive expansion, standardisation, and scientific ratification of Ayurvedic medicine.**

***Keywords*—Ayurvedic medicine, Herbal formulations, Natural language processing, TF-IDF, Ingredient optimization, Clustering, Machine learning, Data-driven Ayurveda.**

I. Introduction

In today's digital age, information has evolved to become the pillar of innovation and decision-making across sectors. Data Science, underpinned by advances in Artificial Intelligence (AI), is enabling researchers and practitioners to derive actionable insights from complex datasets. Of especial significance is the field's potential in the health sector, where accuracy and early treatment make all the difference. Combined applications of machine learning and natural language processing are emerging to provide uncharted avenues of diagnosis, treatment optimization, and drug design [10], [11].

Ayurveda, the ancient Indian system of medicine, has been practiced for thousands of years and is based on principles of balance, natural healing, and individualized treatment. Despite its time-tested legacy and efficacy, Ayurveda has faced challenges in scientific verification and global acceptance due to its qualitative, textual, and tradition-based form of knowledge [1], [9]. With the emergence of data-driven technologies, there now exists an opportunity to bridge this gap by applying modern analytical tools to conventional Ayurvedic data [2], [3].

The purpose of this research is to investigate how Data Science methods can be utilized to refine Ayurvedic drug formulations. With a systematic dataset of Ayurvedic medicine names and ingredients, the study utilizes Natural Language Processing (NLP), clustering algorithms, and visualization techniques to discover common ingredients, classify similar formulations, and identify emerging patterns in the data. This paper concludes by demonstrating that the integration of AI-powered data analytics with traditional medical expertise not only supports standardization and classification efforts but also contributes to the larger goal of mainstreaming Ayurveda—making it both scientifically validated and accessible to global audiences [7], [9].

II. Literature Review

The convergence of Data Science and healthcare has witnessed rapid growth in recent years due to the need for data-driven insights to enhance patient outcomes, support clinical decision-making, and optimize therapeutic interventions. Studies have demonstrated the effectiveness of machine learning and artificial intelligence in medical diagnosis, drug discovery, and personalized treatment

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online)) Volume 14, Issue 1, January 2025)

planning [12], [13]. In ancient systems like Ayurveda, however, the integration of modern analytical methods is only beginning. Ayurveda, a traditional Indian system, is fundamentally different from allopathic medicine, emphasizing personalized treatment through the theory of *Prakriti* (constitution) and diagnostic techniques such as *Trividha Pariksha* (threefold examination). Although Ayurvedic practices are well-established empirically, their formal analysis using computational and statistical approaches remains underexplored. Prior work has stressed the necessity of evidence-based frameworks to achieve global recognition [9]. Subsequent efforts have applied Natural Language Processing (NLP) to classical Ayurvedic texts and clustering techniques to categorize medicinal plants and formulations. However, applying these methods to structured Ayurvedic drug databases for formulation optimization and pattern discovery is still limited. This research addresses that gap using TF-IDF vectorization, K-Means clustering, and PCA visualization on a comprehensive Ayurvedic medicine dataset. Unlike previous efforts relying mainly on text mining or manual classification, this approach offers a fully automated, scalable methodology for clustering and analyzing formulations. It establishes a data-driven framework to translate, categorize, and modernize Ayurvedic knowledge, aiding wider scientific and therapeutic adoption.

III. Methodology

1. Data Preprocessing
   - Missing values were addressed and text data was normalised for consistency. Multi-ingredient strings in the formula descriptions were tokenised with custom regular expressions to handle delimiters such as commas, conjunctions, and special characters.
2. Exploratory Data Analysis (EDA)
   - The most prevalent ingredients were established, including the leading Ayurvedic herbs Ashwagandha and Turmeric. Forms of dosage were analyzed, where tablets and syrups dominated. Manufacturer distributions were plotted to assess the market distribution and diversity of contributions among pharmaceutical companies.
3. NLP and Feature Extraction
   - TF-IDF vectorization was applied to the ingredient data to transform text formulations into numerical vectors. Words that appeared in fewer than two formulations or in more than 85% of records were eliminated to reduce noise. Stopword removal was not done to retain domain-specific words.
4. Clustering
   - KMeans clustering was applied to group similar formulations based on their TF-IDF representations in an attempt to identify inherent thematic patterns in ingredient mixes. Dimensionality reduction via Principal Component Analysis (PCA)

TABLE I.
INTERPRETATION OF K-MEANS CLUSTERS BASED ON AYURVEDIC PRODUCT FEATURES

Cluster	Top Generic Name(s)	Top Dosage Form(s)	Top Brand(s)	Cluster Interpretation
0	Swaschintamani	Capsule	Astro Zest	A niche therapeutic formulation with limited representation. May target specific conditions.
1	Chandanasav, Balarista, Vasakarista, Ashokarista, Mustakarista	Liquid, Capsule, Tablet	Chandanasav, Ashokarista, Balarista	Core Ayurvedic formulations (Asava/Arishta) . Most diverse and widely used therapeutic group.
2	Strophanthus	Capsule	Strophanthus	A rare, possibly cardiotonic product. Represents a highly specific, isolated category
3	Horseradish	Capsule	Elink	Single-ingredient functional product. Likely used for respiratory or digestive issues.
4	Sweet Clover	Tablet	Neurotas	Neurotonic formulation. Specialized product likely targeting cognitive or mental health.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online)) Volume 14, Issue 1, January 2025)

5	Amlapittantak Ras	Tablet, Capsule	Amlapittantak Ras	Formulations targeting acid-peptic disorders. Classical Rasayana products for digestion.
6	Abipattikar Churna	Powder	Abipattikar Churna	Powder-based classical digestive remedy. Distinct group of Churna preparations.

TABLE II

SUMMING UP THE METHODOLOGY

Step	Description	Tools/ Techniques Used	Outcome
1.Data Collection	Loaded the Ayurvedic drug dataset containing names and strengths of medicines.	Pandas	Dataset prepared for analysis.
2.Data Exploration	Explored dataset structure, missing values, and distribution of key fields.	info(), .isnull().sum(), .describe()	Understanding of data completeness and content types.
3.Data Cleaning	Standardised text data by converting to lowercase, removing non-alphabetical characters, and trimming spaces.	String operations in pandas	Cleaned text field ready for feature extraction.
4. Feature Extraction	Transformed drug names into numerical vectors using TF-IDF to capture text patterns	T fidf Vectorizer from sklearn	High-dimensional text feature matrix created.
5. Dimensionality Reduction	Reduced features to 2D space for visualisation.	PCA (Principal Component Analysis)	Simplified cluster visualization.
6. Clustering Analysis	Grouped similar medicines based on textual similarity.	KMeans Clustering	Identified clusters of related drug formulations.
7. Visualization	Displayed top ingredients, clusters, and term distributions.	matplotlib, seaborn, WordCloud	Visual insight into data distribution and groupings.

A. *Visualisation Crux:*

Explaining the Word Cloud and Its Value (Purpose and Analytical Insight)

- Data-Driven Prioritization: The cloud word assists in determining which formulations are most commonly cited, researched, or prescribed. For example, the dominance of terms such as "balarista," "chandanasav," and "rasayan" signifies their foundational status in Ayurvedic practice or literature.
- Trend Detection: It enables researchers to identify trending or under-studied formulations. If names recur less frequently, these could reflect areas of untapped potential for therapeutic investigation or formulation development.
- Visual Knowledge Mapping: Provides an immediate, intuitive snapshot of the scope and concentration of Ayurvedic medicine. It can help researchers and practitioners cope with large datasets.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347 - 6435 (Online)) Volume 14, Issue 1, January 2025)

TABLE III

HOW DATA SCIENCE CAN ACCELERATE GROWTH IN AYURVEDA

Challenge	**Data Science Solution**	**Outcome**
Lack of standardized research	Text mining ancient texts + clinical trials	Evidence-based validation of remedies
Market fragmentation	Clustering manufacturers by product focus	Targeted collaborations and innovations
Ingredient efficacy tracking	Predictive modelling for therapeutic effects	Optimized formulations
Regulatory compliance	DAR number tracking + anomaly detection	Improved quality control
Global demand prediction	Time-series analysis of sales data	Scalable production planning

Embedding data science in Ayurvedic medicine will boost research, quality standardize, and realize global market opportunities through ingredient analysis, predictive modelling, and market intelligence. The convergence of tradition and technology prepares Ayurveda for long-term growth.

IV. CONCLUSION

The application of Data Science in the Ayurvedic medicine sector is a groundbreaking move towards revamping conventional health systems. By conducting this study, we have shown how Natural Language Processing (NLP), clustering, and visualization can extract meaningful insights into Ayurvedic drug formulations. With the application of TF-IDF and KMeans clustering over text data, we were able to classify medicines by composition and gain insights into formulation similarity, ingredient frequency, and optimization areas. This data-intensive method provides a scalable and objective way to categorize and analyse Ayurvedic information, which has been qualitative and textual in nature in the past. Such an analysis not only facilitates standardization of formulations but also improves evidence-based studies, worldwide access, and compatibility with current medicine. Ultimately, the application of Data Science to Ayurveda creates new opportunities for innovation in drug development, personalized therapy, and preservation of knowledge—seamlessly bridging old and new in an effort towards better healthcare.

References

[1] Humber, J. M. (2002). The role of complementary and alternative medicine: Accommodating pluralism. Journal of the American Medical Association, 288, 1655– 1656.

[2] World Health Organization. (2001). Legal status of traditional medicine and complementary/alternative medicine: A worldwide review. Geneva: WHO.

[3] Seidl, P. R. (2002). Pharmaceuticals from natural products: Current trends. Anais da Academia Brasileira de Ciências, 74, 145–150.

[4] ACD Discovery. (n.d.). Retrieved from http://www.ac-discovery.com

[5] Merlion Pharmaceuticals. (n.d.). Retrieved from http://www.merlionpharma.com/index.html

[6] Jiang, Y., Wang, Y., & Yan, X. (2000). Chinese pharmaceutical companies: An emerging industry. Drug Discovery Today, 6, 610–612.

[7] Dubey, N. K., Rajeshkumar, & Tripathi, P. (2004). Global promotion of herbal medicine: India's opportunity. Current Science, 86, 37–41.

[8] Mashelkar, R. A. (1999). Transforming India into the knowledge power. PDRC Report, Government of India.

[9] Patwardhan, B., Chopra, A., & Vaidya, A. D. B. (2003). Herbal remedies and the bias against Ayurveda. Current Science, 84, 1165–1166.

[10] Fabricant, D. S., & Farnsworth, N. R. (2001). The value of plants used in traditional medicine for drug discovery. Environmental Health Perspectives, 109, 69–75.

[11] Maggon, K. (2005). Best selling human medicine 2002–2004. Drug Discovery Today, 11, 739–742.

[12] [12] Ghosh, R., Ghosh, R., & Krishnan, A. (2025). Unveiling the potential of artificial intelligence in revolutionizing diagnosis and treatment of complex diseases: A systematic review. *European Journal of Medical Research*, 30(1), 11.

[13] Dixit, P., Sharma, T., & Singh, A. (2024). Role of artificial intelligence in revolutionizing drug discovery and development. *Current Research in Pharmacology and Drug Discovery*, 5, 100205.