# Machine Learning Methods for Prediction of Biomedical Properties of Nanomaterials

**Nikita Sharma[1], Hajare Hirendra Ramesh[2]**

*[1]Assistant Professor, Department of CS & IT, Kalinga University, Raipur, India.*
*[2]Research Scholar, Department of CS & IT, Kalinga University, Raipur, India.*

Significant advancements have been made in researching the biomedical impacts of nanomaterials, specifically on human health. The study findings published in nanomaterial journals provide overwhelming data, making it difficult to discern the critical research highlights. Using Machine Learning (ML) approaches, automated text mining techniques can easily extract information from a vast collection of texts. The research utilized Naive Bayes and K-means clustering methods on hand-categorized study data sets. The Naive Bayes method achieved a classification result of 89.1% during 5-fold cross-validation on selected libraries. The research used the improved Naive Bayes classification model to anticipate a pattern of research highlights based on the biomedical impacts of Nanomaterials (NM). The analysis included 350,000 original research articles from 22 leading nanomaterial publications spanning 2000 to 2023. Through data mining, polymer NM is the most extensively studied kind of NM. There has been a noticeable decline in research on this material. The study objectives in metallic and carbon-based NM are aligned with those in polymer NM and show an upward trend. The study's emphasis on the prediction of biomedical impacts of NM will mainly revolve around polymers and metallic and carbon-based material structures.

**Keywords:** Machine Learning, Biomedical Properties, Nanomaterials, Prediction.

## 1. Introduction

Nanomedicine, which involves the application of nanoparticles and nanotechnology in the biomedical realm, is a rapidly growing area of study [1]. The Food and Drug Administration (FDA) has only approved a limited number of nanoparticle systems. There is a solid need to expedite the process of applying nanoscience findings from the laboratory to clinical use. Although there are well-established experimental techniques in nanomedicine studies, computational support for nanomedicine is less advanced [2]. There is a significant need for

more reliable data sources that non-informatics professionals can easily access. Utilizing Quantitative Structural Activity Relations (QSAR) approaches, specifically in nanomedicine, such as "nano-QSARs" and other predicting designs, can expedite the translational procedure in the area of nanomedicine [14]. Non-informatics experts interested in nanomedicine need help accessing data on fascinating nano-QSAR methods. The objectives of this research are to assess the research on the utilization of data mining and Machine Learning (ML) for predicting biomedical characteristics of medically relevant Nanomaterial (NM) and to analyze the advancements and obstacles that the emerging field of nano-QSAR technique encounters in its efforts to contribute to the design of efficient nanomedicines significantly [4].

ML is a nascent division of Artificial Intelligence (AI) [5]. AlphaGo, an AI built by Google, has been consistently winning against the world championship of Go in recent times. This demonstrates the significant potential of ML in several aspects of people's everyday lives. ML has been extensively used in other technological domains, including computer vision, mining data, robotics, and medical diagnostics, resulting in significant accomplishments. ML is the field that examines how computers imitate or execute human learning patterns to gain new information or abilities and rearrange existing knowledge to enhance their capabilities consistently [3]. These methods can identify and extract data characteristics from large datasets and uncover the underlying patterns and rules governing the information [6]. In the current era of information, large amounts of data are being created from literary sources. As a result, distinct ML techniques have been developed for text-mining tasks [11]. In such applications, ML methods are used to acquire the ability to do a particular job. Once the models have learned the necessary rules, they utilize this knowledge to make judgments or predictions regarding real-world occurrences. This ML technique makes it a potent tool in text-mining scenarios. The proliferation of educational studies has led to a significant rise in research information and publications [9].

The research on the biomedical impacts of nanomaterials is a crucial and significant area of research on the present frontier of scientific inquiry, with direct implications for human welfare. To examine the relationship between novel NM and biomedical processes, it is possible to diminish the harmful effects of these developing materials on living organisms. One leverages their distinctive physical and chemical characteristics to create NM that benefits humans. To assist scholars in navigating the study materials on the biomedical impacts of NM, there have been dedicated initiatives towards data mining in biomedical studies [12].

## 2. Materials and Methods

### 2.1 Source of data

The research has selected six groups to classify typical nanomaterials: "Carbon NM," "Ceramic NM," "Metallic NM," "Organic NM," "Polymer NM," and "Semi-metallic NM." The PubMed search engine retrieves research papers with titles and abstracts of six standard NM classes [7]. This uses specific search phrases and the "Best Match" screening parameters. For further information on the search phrases and associated choices. The

research eliminated the redundant documents in both classes. The overall number of papers acquired is 12000 articles. This includes 3300 articles on carbon NM, 1000 pieces on ceramic NM, 3000 articles on metallic NM, 1200 articles on organic NM, and 2000 publications on polymer NM. One thousand thirty items have properties of both metals and semiconductors.

The research conducted a research analysis to forecast the emerging trends in the biomedical impacts of NM. This analysis was based on the findings from 22 leading NM journals, such as 350,000 original scientific articles published between 2000 and 2023. The research used ML to classify data analysis and predict research trends.

2.2. Text representation

The data included in the original paper is presented in plain language [8]. ML systems cannot process these messages directly. The research must create a mapping between the initial text dataset and the text representation approach and then utilize that framework as the input for the method.

The paper preprocessing approach involves the segmentation of words in phrases using spaces, as well as the elimination of punctuation and stop phrases that do not add to semantic comprehension. The primary textual representation methods include the Boolean approach, vector space approach, topic design, and noise-reducing automated encoder design. These methods are chosen based on specific needs. The preprocessing was executed using a customized Python script that includes the removal of punctuation from every file (except the English hyphen), eliminating stop phrases, and converting to lowercase characters.

2.3. Feature dimension reduction

Vector space models encode texts using high-dimensional and sparsely populated characteristics [13]. The utilization of high latitude vector spaces decreases the computational performance and leads to the classification performing poorly in classifying fresh data due to over-fitting. The specific terms have distinct significance in text categorization. Filtering the attribute words and picking the characteristic items with significant content enhances the classification efficiency of the system. The dominant strategies for reducing feature dimensions are broadly classified into two main groups: choosing and extracting features.

The textual representation employs a vector-based model that relies on the frequency of words. The initial dimension of the space before reducing its size is 38000. The feature choice is conducted using the word frequency length sorting approach, and only the feature items with a frequency ranking inside the top 3000 are chosen. The whole spatial dimension is decreased to 3000 units.

2.4. Establish a classification model

Clustering and categorization are primary objectives in the field of data mining [10]. The suggested clustering analysis addresses the issue of processing unlabeled information using a method. A preliminary relevant text clustering algorithm has been developed, which is utilized to group unlabeled content and extract information using current techniques. This investigation used conventional clustering techniques such as the K-means and Naïve Bayes

techniques. The K-means technique is a very efficient clustering approach used in data mining to analyze clusters. The method is enhanced by optimizing its initial center choices, choosing features, and automating the determination of clustering amount and length for outlier elimination. The Naive Bayes categorization method relies on the premise of attribute independence, meaning that it assumes the distinctive characteristics are not influenced by each other.

2.5. Performance evaluation

The conventional metrics for classification assessment assess the classifier's effectiveness and provide insights into the particular implications of the findings. The confusion matrix clearly represents the relationship between the anticipated category and the actual categorization of the specimens. It illustrates the distribution of the outcomes and indicates the classifier's identification efficiency. Cluster reliability refers to assessing the impact of clustering in clustering research. Purity measures the percentage of articles in a cluster that belong to a specific category relative to the overall amount of articles in the clusters. It is considered the greatest when it aligns with a recognized category.

## 3. Results

The Region of Convergence (ROC) curve was generated by computing the likelihood of a test sample being assigned to a particular group using the Naive Bayes approach in the categorization process. In the classifying procedure, the Naive Bayes technique is used to compute the likelihood of the test sample falling into a particular group. This likelihood is then used to display the ROC curve. Figure 1 shows the ROC curves of six representative NM individually. The classification is more effective when the curve is closer to the top left corner. Three of these groups are distinct. The polymer NM achieved the largest Area Under the Curve (AUC), reaching a value of 0.9289. The findings demonstrate that the Bayesian classification can accurately classify biomedical impact studies of predicting nanoparticles.

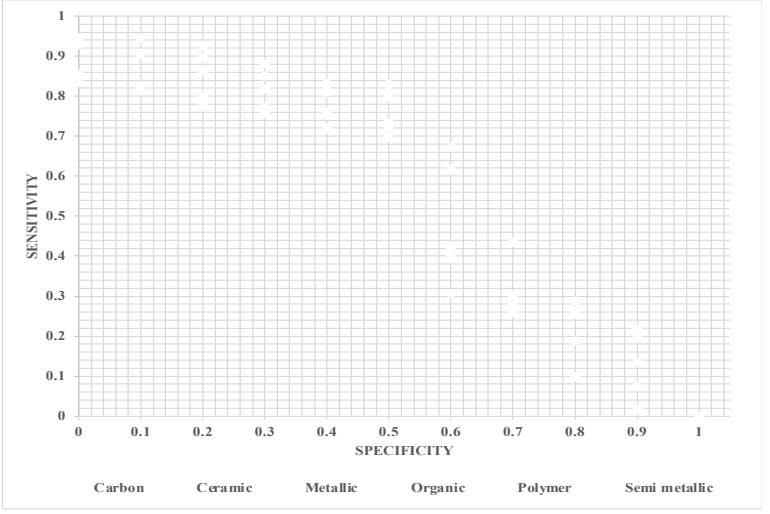


Fig. 1. ROC curve analysis of different Ml models

Biomedical materials have a clear and significant edge in the number of publications connected to polymers. Biomedical materials are the primary focus of biomedical investigations, as described in the scope of the magazine's study. The Journal of the American Chemical Academy and Langmuir are both focused on studying the biomedical impacts of nanotechnology, specifically about typical NM. This research is based on data extraction from the top 22 journal classifiers. Among the six types of NM studied, the top three are polymers, metallic, and carbon-based NM.

In NM systems, polymer NM has been the subject of many studies. The number of journals related to this topic has decreased. The popularity of metallic and carbon NM has been increasing steadily, with the former being ranked as the second most studied material and the latter as the third most studied material. The current research trend centers on studying polymers and metallic and carbon NM, particularly their biomedical effects. This observation is based on an analysis of research on NM investigations. Out of the 22 leading publications on NM, more papers are needed to focus on ceramic and semi-metallic NM.

## 4. Conclusion

Studying the biomedical effects of NM is a crucial and significant scientific topic that plays a fundamental role in human health. It has yielded substantial research accomplishments. To assist scholars in navigating the vast and rapidly expanding field of NM's biomedical impacts, the research employed ML techniques to conduct data mining on the findings. Through this process, the research was able to identify and analyze new study hotspots and predict and identify emerging trends in this area.

The research developed a very efficient NM classification using the Naive Bayes method and the K-means method, achieving an accuracy of 89.1%. The research utilized a highly efficient classification model to forecast study patterns and focal points on the biomedical impact of NM, drawing from 22 state-of-the-art NM-focused journals. The polymer NM is now the most extensively explored material due to its inclusion of several types of conventional biomedical materials. While the study and development of medicinal effects have increased, investigations on typical polymer NM have declined over the years. They remain among the top three most popular NM. The research on metallic NM is rapidly increasing to match the research on polymer NM. The study focused on carbon NM, which saw a modest rise and remained consistently high. The study's emphasis on investigating the biomedical effects of NM will primarily be on polymer, metallic, and carbon NM.

## References

1. Zhang, C., Yan, L., Wang, X., Zhu, S., Chen, C., Gu, Z., & Zhao, Y. (2020). Progress, challenges, and future of nanomedicine. Nano Today, 35, 101008.
2. Jiang, W., Wang, Y., Wargo, J. A., Lang, F. F., & Kim, B. Y. (2021). Considerations for designing preclinical cancer immune nanomedicine studies. Nature nanotechnology, 16(1), 6-15.
3. Srinivasa Rao, M., Praveen Kumar, S., & Srinivasa Rao, K. (2023). Classification of Medical Plants Based on Hybridization of Machine Learning Algorithms. Indian Journal of Information

Sources and Services, 13(2), 14–21.
4. Konstantopoulos, G., Koumoulos, E. P., & Charitidis, C. A. (2022). Digital innovation enabled nanomaterial manufacturing; machine learning strategies and green perspectives. Nanomaterials, 12(15), 2646.
5. Singh, A. V., Rosenkranz, D., Ansari, M. H. D., Singh, R., Kanase, A., Singh, S. P., ... & Luch, A. (2020). Artificial intelligence and machine learning empower advanced biomedical material design to toxicity prediction. Advanced Intelligent Systems, 2(12), 2000084.
6. Rosa, C., Wayky, A.L.N., Jesús, M.V., Carlos, M.A.S., Alcides, M.O., & César, A.F.T. (2024). Integrating Novel Machine Learning for Big Data Analytics and IoT Technology in Intelligent Database Management Systems. Journal of Internet Services and Information Security, 14(1), 206-218.
7. Salem, S. S., Hammad, E. N., Mohamed, A. A., & El-Dougdoug, W. (2022). A comprehensive review of nanomaterials: Types, synthesis, characterization, and applications. Biointerface Res. Appl. Chem, 13(1), 41.
8. El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. Expert systems with applications, 165, 113679.
9. Mojail, N. Disages K., et al. "Understanding Capacitance and Inductance in Antennas." National Journal of Antennas and Propagation 4.2 (2022): 41-48.
10. Afnaan, K., Peeta, B.P., Tripty, S., & Bhanu, P.K.N. (2014). Comparative Analysis for Feature Extraction and Prediction of CKD Using Machine Learning. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), 15(2), 202-225. https://doi.org/10.58346/JOWUA.2024.I2.014
11. Ahmed, S. T., Sreedhar Kumar, S., Anusha, B., Bhumika, P., Gunashree, M., & Ishwarya, B. (2020). A generalized study on data mining and clustering algorithms. New Trends in Computational Vision and Bio-inspired Computing: Selected works presented at the ICCVBIC 2018, Coimbatore, India, 1121-1129.
12. Shamsudin, N. F., Ahmed, Q. U., Mahmood, S., Ali Shah, S. A., Khatib, A., Mukhtar, S., ... & Zakaria, Z. A. (2022). Antibacterial effects of flavonoids and their structure-activity relationship study: A comparative interpretation. Molecules, 27(4), 1149.
13. Bobir, A.O., Askariy, M., Otabek, Y.Y., Nodir, R.K., Rakhima, A., Zukhra, Z.Y., Sherzod, A.A. (2024). Utilizing Deep Learning and the Internet of Things to Monitor the Health of Aquatic Ecosystems to Conserve Biodiversity. Natural and Engineering Sciences, 9(1), 72-83.
14. Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: a review. Complex & Intelligent Systems, 8(3), 2663-2693.
15. Kononova, O., He, T., Huo, H., Trewartha, A., Olivetti, E. A., & Ceder, G. (2021). Opportunities and challenges of text mining in materials research. Iscience, 24(3).