Reviews on Advanced Materials and Technologies, 2025, vol. 7, no. 4, pp. 203–210
DOI: 10.17586/2687-0568-2025-7-4-203-210

Development of an AI Tool for Systematic Review of Scientific Articles Containing IR-Spectra of Boron Nitride Nano- and Microparticles

I.M. Sosnin, A.R. Reznikova, A.V. Frolova

Laboratory of Chemical Methods for Materials Elemental Analysis, Togliatti State University, Belorusskaya str. 14, Togliatti, 445020, Russia

Article history

Received November 16, 2025
Accepted November 30, 2025
Available online December 03, 2025

Abstract

In this work a creation of neural network method and its application to exploration of optical-property data of hexagonal boron nitride nano- and microparticles is presented. In particular, the method analyses the data of electromagnetic absorption in the infrared region. The work shows how modern algorithms of natural language processing and deep-learning can be used for automatization of search and analysis of raw data. In the work we apply deep neural network models including convolutional neural network (CNN) for review of infrared spectra and transformers (SciBERT, ChemBERTa) for examination of text information. Multimodal learning, integrating CNN and semantic analysis of texts, was developed for survey heterogeneous data.

Keywords: IR-spectrum; Boron nitride; RAW Data; CNN; ChemBERT

1. INTRODUCTION

The modern science faces exponential growth of volume of scientific data. It can be systematized with neural networks. These technologies have progressed leaps from 1950s, when Alan Turing presented the first investigation on this subject [1]. To date, natural language processing (NLP) has paved the way toward automation of discovery and harmonisation of research data. For instance, He et al. [2] demonstrated a neural network, that sheds light on role of each raw component, utilized in chemical synthesis of inorganic substances, integrating literature published earlier. In another paper He et al. [3] described a predictive model for synthesis conditions, leveraging intelligent search of published papers. The neural network has screened papers documenting the synthesis of target substances. After that it was tasked with predicting the list of required precursors, which was documented in withheld article. he prediction was considered successful if at least one precursor from the list was mentioned in the withheld article. Model output exhibited 82% experimental accuracy. Moreover, neural network-based predictions of material properties trained on publication data appear in the literature with experimental verification. Tshitoyan et al. [4] presented a model that predicted the existence of materials with desired thermoelectric properties.

However, in certain cases modelling exhibits ambiguous results. For instance, predicting the shape of the absorption spectrum in the infrared region for nano- and microparticles having different morphologies is a formally unparameterizable problem. The presence of correlation between a nanoparticle morphology and its optical properties in the infrared region can be inferred from theoretical principles, published by Halford [5] in 1946. He attributed the shape of absorption spectrum to external form of a crystal within the framework of his theory of site groups. Since that time, the theory has been evolving dynamically. Currently, these ideas are gaining experimental

* Corresponding author: I.M. Sosnin, e-mail: i.sosnin@tltsu.ru

support. For example, in the article [6] a simple method based on Fourier transform infrared spectroscopy (FTIR) is shown enabling quantification of nanoplate fraction in powder of boron nitride nanotubes. It is known that a FTIR-spectrum of boron nitride has two peaks. However, position of each peak differs across publications. The first one has a value of 697 cm^{-1} [7], 805 cm^{-1} [8], 809 cm^{-1} [9]; the second—1365 cm^{-1} [10], 1379 cm^{-1} [8], 1387 cm^{-1} [9], 1400 cm^{-1} [11], 1401 cm^{-1} [7], 1634 cm^{-1} [7]. As can be seen, these values exhibit considerable variation which is likely associated with size and morphology of nanoparticles. Comparing FTIR-spectrum and results of scanning electron microscopy (SEM) can reveal a correlation between a nanoparticle morphology and its optical properties in the infrared region. It is obvious that such a challenge can be efficiently overcome with artificial intelligence (AI).

Development of an AI-system is complicated due to high data heterogeneity, since publications include different types of RAW-files. Furthermore, the AI-system should label datasets to include RAW-files for subsequent annotation. Then it should interpret the found results. Using hybrid models integrating deep learning, NLP, and multimodal methods is necessary to address these limitations. For annotating of labeled data, the pretrained SciBERT model within the Hugging Face Transformers framework is employed, enabling context-aware analysis through transformer-based learning mechanisms [12]. ChemDataExtractor serves as a specialized tool for automated extraction of structured chemical-compound data from scientific texts. Beyond identifying molecular structures, the system can recognize and extract numerical parameters, experimental characteristics, and procedural conditions. RDKit provides functionality for manipulating chemical structures, while the Semantic Scholar API grants access to extensive collections of scientific publications [13].

In the present study various models and their application for semantical analysis of scientific texts are considered, and a search algorithm for academic publications is offered. Built on aforementioned models an original method is proposed to address the challenges outlined above. The method performance is demonstrated via search and analysis of the scientific articles containing RAW data files on optical properties of hexagonal boron nitride nano- and microparticles various shapes.

2. RELATED WORKS

Modern approaches to automated analysis of scientific publications increasingly rely on deep learning and neural architectures to process heterogeneous data types such as texts, images, and a structured content. Accordingly, integration of transformer-based language models and convolutional neural networks (CNNs) has proven especially effective for extracting meaningful patterns from complex scientific data [14]. However, creation of hybrid models is elusive with a dilemma: models are typically either good at uncovering deep, latent patterns or at providing transparent explanations for simple ones—but rarely excel at both. Designing hybrid systems that balance these objectives remains difficult. Our work focuses explicitly on the search and discovery aspect of this challenge, rather than explainability, emphasizing the capacity of the system to detect subtle and multifaceted relationships in scientific data.

To achieve the aforementioned goal, we build on recent advances in domain-adapted transformer models. SciBERT [15], a variant of the general-purpose BERT model [16], is specifically trained on large-scale scientific corpora and has demonstrated improved performance on scientific NLP tasks. It offers robust contextualized embeddings for research-specific terminology and sentence structures, which are essential for understanding publication content. In our architecture, SciBERT is used as the primary mechanism for document representation, capturing the semantics of scientific abstracts and full texts. Concurrently, for multimodal data types such as experimental spectra, we employ classical CNNs. CNNs remain one of the foundational methods in pattern recognition, particularly effective in capturing local feature hierarchies in image-like or signal-based data [17]. The original application of CNNs to visual tasks, such as handwritten digit recognition, laid the groundwork for their later success in scientific domains involving time series, spectral data, or microscopy images. In our system, CNNs are used in conjunction with autoencoders to reduce dimensionality and extract compact latent features from raw scientific measurements. By combining SciBERT and CNNs within a unified architecture, our model demonstrates a high capacity for uncovering latent patterns across both textual and numerical domains. This emphasis on search-oriented intelligence distinguishes our work from others that prioritize model interpretability. Rather than simplifying insights for human consumption, we focus on maximizing discovery potential, thereby providing researchers with new connections and hypotheses drawn from complex, multimodal datasets.

3. METHODOLOGY

A comparative analysis was conducted to evaluate the efficiency of various neural network architectures, including BERT, SciBERT, and ChemBERTa. These models are specifically designed for scientific texts and enable deep semantic analysis; the results are presented in the Table 1.

The Table 1 highlights crucial characteristics of three popular transformer-based language models for NLP, with a focus on scientific and chemical domains. The

Table 1. Comparative analysis of the efficiency of different neural network architectures.

Parameter	BERT	SciBERT	ChemBERTa
Architecture	Transformer	BERT-based	BERT-based
Training Corpus	General corpus	1.14M articles	Chemical texts
Application	General NLP tasks	Scientific text analysis	Chemistry
Accuracy	Moderate	High (scientific data)	High (chemical data)
Generalization	Good for general language	Strong for scientific articles	Strong in chemistry, weaker on general texts
Training	Pretrained on general corpus	Domain-specific pretraining	Specialized chemical pretraining
Industry use	NLP apps, chatbots	Scientific research	Chemical & pharma

analysis reveals a clear trend in the evolution of language models from general-purpose to highly specialized solutions. Results indicate that base BERT significantly underperforms compared to domain-specific counterparts when handling professional scientific and chemical terminology despite good throughput. Beltagy et al. [12] show that SciBERT achieves accuracy of processing academic texts by 15–30% higher due to being pretrained on scientific articles. This supports the hypothesis that language model effectiveness in specialized domains directly correlates with the alignment between pretraining data and the target domain. However, increased specialization inevitably reduces a model ability to generalize across related domains, necessitating careful architecture selection based on specific use cases. Practical experience suggests SciBERT is optimal for interdisciplinary research, while ChemBERTa is preferable for highly specialized chemical tasks.

4. EXPERIMENT

The developed algorithm is a comprehensive system for automated screening of scientific publications, implemented in Python using modern machine learning techniques. The system architecture, illustrated in Fig. 1, consists of two core data processing stages, each implemented as separate modules integrated into a unified analytical framework. First, data format detection is performed using automatic recognition algorithms that analyse file signatures and their internal data structures. Second, a multimodal neural network architecture processes text, images, and numerical data simultaneously.

Users can initiate the analysis through three ways: by entering keywords for publication searches, uploading PDF files from a local device, or sending direct API requests to specialized databases. During the data retrieval phase, the system queries three major scientific sources in parallel. Access to Google Scholar is facilitated via the scholarly library, while Semantic Scholar and PubMed are accessed through their official APIs. This stage is highlighted as a separate zone in the diagram, emphasizing its autonomy within the overall workflow. The data extraction process involves universal handling of heterogeneous content. A module based on PyMuPDF analyses PDF document structures, extracting both textual and graphical elements. HTML parsing via BeautifulSoup enables the

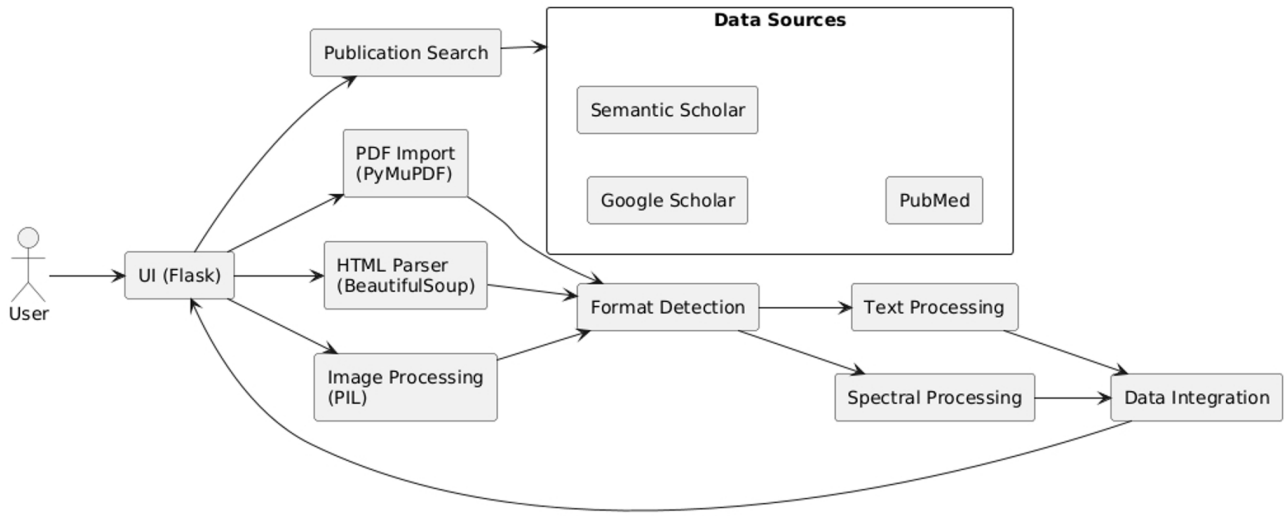

Fig. 1. Workflow diagram of developed neural network algorithm.

detection and download of raw data-tables with experimental measurements, spectra files, and supplementary materials. A dedicated component processes images using the PIL library. A crucial feature of the system, represented by a diamond-shaped branching element in the diagram, is the separate processing of different data types. Textual content undergoes multiple transformation stages. The SciBERT model generates document embeddings, which are then clustered by topic using a combination of TF-IDF and the K-means algorithm. For spectral data, a fundamentally different transformation pipeline is applied—feature extraction via the 1D-CNN followed by dimensionality reduction using an autoencoder. The integration phase synthesizes heterogeneous data, establishing correlations between textual descriptions and experimental spectra. The final step involves result visualization through a Flask-based web interface, where users access the analysis outcomes.

The system supports export of the results in JSON, CSV, and PNG formats and includes functionality for automatic raw data retrieval via HTML structure analysis. Implementing all modules within a unified Python architecture ensures seamless integration into existing research workflows and scalability for similar tasks in materials science.

5. RESULTS

The developed system takes the form of complex web application, created and powered by Flask, which delivers multifunctional analysis of scientific publications. At startup the system initiates a web server and offers user an interface for entering search queries and uploading files of various formats. The functionality consists of two data processing stages: search and publication processing. For implementation of scientific publication search, the system is integrated with three bibliographic databases: Google Scholar (via scholarly), Semantic Scholar (via official API) and PubMed (via E-utils API). Each source is processed with a specific module, which perform normalization of obtained metadata, including publication title, abstract, publication year and citation index. Particular attention is given to processing of failure mode during screening the Google Scholar without official API. Under these conditions the request can be blocked so that the system implements a repeated attempt.

The system supports multiple file formats: PDF-documents, CSV-tables and images of spectra. Structure of the system includes three neural networks: a one-dimensional convolutional neural network (1D CNN) for screening of spectral data, an autoencoder for compression and extraction of spectra and SciBERT model for processing of scientific texts. These models were built by integrating Keras and TensorFlow libraries so they provide information extraction, semantic analysis and data preprocessing for clustering. An architectural feature of the system is the usage of a unified HTML template so that the application becomes self-sufficient and easy to deploy. Output for a user is a structural data of publications.

The system shows high speed of data processing. Search of publications by key words is performed within 10–15 seconds for 5 requests. Extraction of text from PDF files required 2–5 seconds per page, while characterization of 100 publications with K-means was performed within 3–7 seconds. SciBERT delivers up to 92.4% accuracy, CNN displays accuracy of 85–90% for spectra analysis. Input data was labeled with the 1D-CNN and then it was annotated with SciBERT. The network architecture consists of a convolutional layer with 32 filters of size 3, followed by a max-pooling operation with a stride of 2. The output is passed to a fully connected layer that maps the data into a 64-dimensional feature space. During training, the Adam optimizer was used with a learning rate of 0.001 and mean squared error (MSE) as the loss function. Evaluation on a test set of boron nitride spectrum demonstrated a classification accuracy of 89.2% for particle morphology types. The data autoencoder of spectrum follows a symmetric encoder-decoder design. The encoder transforms the input through two fully connected layers with 64 and 32 neurons, respectively, using ReLU activation in the first layer and a linear function in the second. The decoder mirrors this structure, performing the inverse transformation and ending with a sigmoid activation in the output layer. Batch normalization was added between layers to accelerate training, reducing convergence time by 37% compared to the baseline architecture. The model was trained on a large dataset of 12,000 spectra from the NIST database, achieving a compression ratio of 3.125:1 while maintaining reconstruction quality, with a peak signal-to-noise ratio (PSNR) of 28.4 dB.

A modified version of the SciBERT model with AllenAI is used for processing textual information. This pretrained architecture, consisting of 110 million parameters and 12 attention layers, was further fine-tuned on a corpus of 45,000 scientific articles in the field of materials science. The model employs WordPiece tokenization with a vocabulary of 30,000 tokens and can process text segments up to 512 tokens in length, converting them into 768-dimensional vector representations. In scientific topic classification tasks, the model achieves an F1 score of 0.924, while the accuracy of chemical formula extraction reaches 89.1%. A comparative analysis of computational performance shows that the 1D-CNN processes a single spectrum in an average of 8.2 ms, the autoencoder in 5.7 ms, whereas text processing with SciBERT takes approximately 142 ms. Memory consumption varies significantly across models, ranging from 2.1 MB for the autoencoder to 438 MB for the SciBERT. This resource

Table 2. Comparative performance characteristics of existing approaches to processing scientific information.

Criteria	Manual input	DeepSeek	The present system
Duration of processing	15–30 minutes	2–5 minutes	1–2 minutes
The number of sources	1–2	1	3 parallel
Duration of search of supplementary files	About 1 hour	–	30–60 seconds
Duration of PDF file processing	10–15 minutes per a file	2–3 minutes	5–10 seconds
Natural language processing	Cursory assessments	Deep search	SciBERT
Clustering of results	Manual	Set up	Automatically

distribution enables efficient scaling of the system for processing large volumes of scientific data.

All models are integrated into a unified analytical pipeline: the SciBERT first processes the textual metadata, followed with the 1D-CNN, which analyzes spectral data, and finally the autoencoder generates compact representations for subsequent clustering. The implementation is based on PyTorch 1.12 with CUDA 11.3 support, enabling an average processing time of approximately 0.87 seconds per publication, including all associated data, when running on an NVIDIA T4 GPU. To illustrate the advantages of our approach, the Table 2 presents a comparative analysis of three strategies for working with scientific information, namely traditional manual search by researchers, general-purpose neural network solutions, and the proposed algorithm.

Based on the obtained results, it can be concluded that the algorithm combines high processing speed with a comprehensive approach to scientific information analysis. The advantages are particularly evident when handling large datasets—conducting systematic reviews, analyzing new research directions, or searching for experimental data. The solution not only saves up to 90% of researchers' time but also delivers more complete and systematic results compared to alternative methods.

For document analysis, the upload PDF function extracts the first 3000 text characters and all images, displaying formatted text and image counts. Required libraries include PyTorch, Transformers, scholarly and web application framework Flask. Stable internet connectivity is essential for API communication, and users must have read/write permissions for the upload directory. To begin, users execute the Python script, launching a web interface at http://localhost:5000. The main page offers two core functions: article searches by user-specified topics and PDF uploads for text/image extraction. For publication searches, entering keywords and clicking "Search" automatically queries all three databases. Results are displayed as an article list with titles, sources, publication years, and citation counts. An optional "Find related files" checkbox activates PDF/CSV/image retrieval, with links displayed as shown in Fig. 2. For analysis of downloaded documents, a PDF file upload feature has been implemented. After clicking the "Upload PDF" button and selecting the file, the algorithm extracts the text content (the first 3000 characters) and all the images contained in the document. The processing results will be displayed on the page: the text will be shown as a formatted block, and the number of extracted images will be indicated separately.

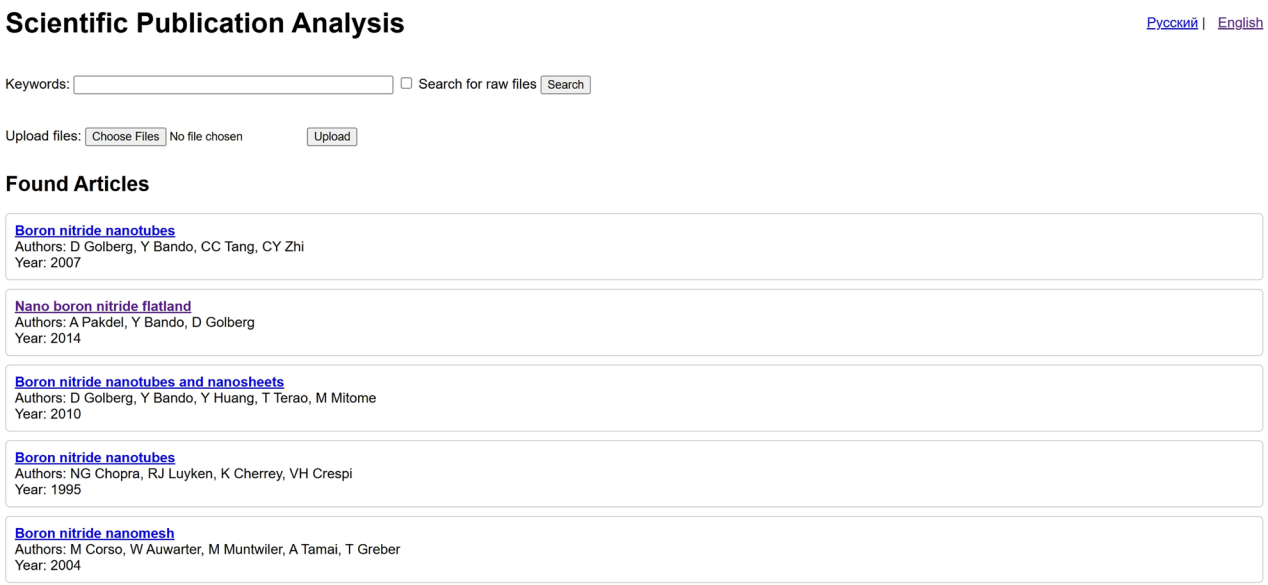

Fig. 2. Algorithm output example.

6. DISCUSSION

The code implementing the basic search functionality is presented. Evaluation results demonstrated high applicability of the specified models for automating scientific document processing and extracting valuable information. Our team developed the method of integrating chemical compound and database analysis results. Several improvements are proposed to enhance the algorithm efficiency. One of the main development directions involves adding support for additional scientific publication databases. This will expand the range of available information sources. To improve algorithm reliability, we propose implementing error handling, including request retries when failures occur. This will prevent issues related to temporary Google Scholar unavailability or possible query limitations. Additionally, we plan to modify the results presentation format to enable data export as JSON or CSV files. The proposed algorithm will serve as the foundation for a web application that will provide convenient access to scientific literature, making the publication search and analysis process more efficient. Such an application could be valuable for researchers, educators, and students who require quick and convenient access to up-to-date scientific information. Further development plans include creating a web application based on the proposed algorithm that will allow users to search scientific publications through a user-friendly interface.

The developed system is built upon a set of modern neural network architectures specifically adapted for processing scientific texts and chemical data. The key component of the system is SciBERT—a specialized version of the BERT model pretrained on a corpus of extensive scientific publication. This model was selected due to its unique capability to understand complex scientific terminology through training on academic texts, fundamentally distinguishing it from publicly available language models. SciBERT provides contextual word representations, enabling identification of semantic relationships even in the absence of exact lexical matches, which is critically important for working with scientific literature. The model also supports fine-tuning for specific tasks such as publication classification or chemical compound extraction from text. For processing lengthy scientific articles, the system additionally employs specialized architectures like Longformer and SPECTER, optimized for working with long documents. These models overcome traditional transformer context-length limits while maintaining high analysis accuracy. CNNs also play several key roles in scientific data processing. Primarily, CNNs are used for analyzing visual content in scientific publications, proving particularly valuable when working with two-dimensional molecular structure representations. For processing tabular data from scientific publications, we utilize one-dimensional convolutional networks (1D-CNN) that effectively identify local patterns in numerical sequences. Promising research area for further system development includes replacing the 1D-CNN with more advanced architectures such as WaveNet for improved spectral peak extraction, applying knowledge distillation techniques to reduce the size of the text model, and incorporating attention mechanisms into the autoencoder to enhance the interpretability of results.

Data collected with our AI-tool can be analyzed through the theoretical framework of interaction between infrared radiation and a crystal. Starting from Frenkel's work [18] and his analogy between a crystal and molecule we can treat the IR spectrum as a collective characteristic of all chemical bonds in a particle. The representation of optical properties of B–B and B–N bonds can be guided by the postulated framework of unpaired electron interactions, as proposed by Pauling [19] and the electronegativity scale he established, later refined by Oganov and Tantardini [20]. The correlation between a crystal shape and its IR absorption, published by Halford [5], provides a theoretical background for interpreting spectral data. He showed that each unit cell includes several variants of molecule arrangement into the unit cell, which are called site groups. Optical effects on crystal surface can be predicted using site group theory. Given that hexagonal boron nitride (h-BN) has the $3D_{3h}(2)$ site group, we can use Halford's correlation to infer particle shape distribution from IR spectral data. Beyond that correlation approach postulated by Fateley et al. [21] provides a method for identifying IR-active lattice vibrations and analysis surfaces elements of symmetry. The resulting theoretical framework allows for comprehensive analysis of the collected spectral data.

7. CONCLUSION

Application of the developed neural network method significantly simplifies the task of searching, analysing, and systematizing scientific articles, particularly when handling RAW files. The proposed approach has demonstrated high efficiency in addressing current challenges in scientific data analysis. The system successfully overcomes the issue of high-data format heterogeneity through its built-in automated file type classification system [22]. The lack of labelled datasets is addressed via preliminary clustering and semantic filtering using transformers [23]. The interpretability challenge is partially resolved through textual annotations with user-adjustable features [24]. Furthermore, the neural network supports both Russian and English texts, making it a universal tool for international research tasks.

Future advancements in this field involve developing multimodal models and establishing RAW data rep-

resentation standards, which will significantly improve the quality of analysis for natural science publications. Neural network technologies for scientific publications continue to evolve. Autoencoders were additionally employed to reduce data dimensionality and identify latent patterns. The final product is an intelligent scientific publication analysis system equipped with Explainable AI (XAI) mechanisms to ensure full algorithm transparency [25]. The developed solution enables researchers to not only locate relevant publications in leading scientific databases but also obtain detailed explanations of the system operation principles and factors influencing search results. The Grad-CAM mechanism analyses spectral data, identifying which specific spectral regions had the greatest impact on classification. To explain publication trends, the algorithm automatically detects crucial events and scientific breakthroughs that affected publication volumes in specific years. An interactive web interface of the system comprises three main sections: a search panel with keyword input capability, a query history tab storing the last 10 search operations, and a help section with detailed instructions. The system proves particularly effective for conducting systematic literature reviews, analysing research trends, and identifying promising scientific directions. This solution combines the high performance of automated analysis with the transparency and interpretability required for rigorous scientific work, making it suitable for both individual research and large-scale academic projects.

ACKNOWLEDGEMENTS

This research was funded by the Russian Science Foundation grant # 25-23-00477 “Investigation of the influence of morphology of nano- and microparticles of boron nitride on mechanical properties of magnesium-matrix composites”.

REFERENCES

[1] A.M. Turing, I.—Computing machinery and intelligence, *Mind*, 1950, vol. LIX, no. 236, pp. 433–460.

[2] T. He, W. Sun, H. Huo, O. Kononova, Z. Rong, V. Tshitoyan, T. Botari, G. Ceder, Similarity of precursors in solid-state synthesis as text-mined from scientific literature, *Chemistry of Materials*, 2020, vol. 32, no. 18, pp. 7861–7873.

[3] T. He, H. Huo, C.J. Bartel, Z. Wang, K. Cruse, G. Ceder, Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature, *Science Advances*, 2023, vol. 9, no. 33, art. no. eadg8180.

[4] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K.A. Persson, G. Ceder, A. Jain, Unsupervised word embeddings capture latent knowledge from materials science literature, *Nature*, 2019, vol. 571, pp. 95–98.

[5] R.S. Halford, Motions of molecules in condensed systems: I. Selection rules, relative Intensities, and orientation effects for Raman and infra-red spectra, *The Journal of Chemical Physics*, 1946, vol. 14, no. 1, pp. 8–15.

[6] H. Harrison, J.T. Lamb, K.S. Nowlin, A.J. Guenthner, K.B. Ghiassi, A.D. Kelkar, J.R. Alston, Quantification of hexagonal boron nitride impurities in boron nitride nanotubes via FTIR spectroscopy, *Nanoscale Advances*, 2019, vol. 1, no. 5, pp. 1693–1701.

[7] A.P. Bhasi, N. Hanna Wilson, T. Palanisamy, Nanosized hexagonal boron nitride and polyethylene glycol-filled leathers for applications demanding high thermal insulation and impact resistance, *ACS Omega*, 2022, vol. 7, no. 50, pp. 45120–45128.

[8] L. Xu, Y. Peng, Z. Meng, W. Yu, S. Zhang, X. Liu, Y. Qian, A co-pyrolysis method to boron nitride nanotubes at relative low temperature, *Chemistry of Materials*, 2003, vol. 15, no. 13, pp. 2675–2680.

[9] J.-Q. Hu, Q.-Y. Lu, K.-B. Tang, S.-H. Yu, Y.-T. Qian, G.-E. Zhou, X.-M. Liu, J.-X. Wu, Synthesis and characterization of nanocrystalline boron nitride, *Journal of Solid State Chemistry*, 1999, vol. 148, no. 2, pp. 325–328.

[10] Y. Wang, K. Zhang, L. Ding, L. Wu, E. Songfeng, Q. He, N. Wang, H. Zuo, Z. Zhou, F. Ding, Y. Hu, J. Zhang, Y. Yao, An efficient boron source activation strategy for the low-temperature synthesis of boron nitride nanotubes, *Nano-Micro Letters*, 2025, vol. 17, art. no. 25.

[11] C. Tang, Y. Bando, Y. Huang, C. Zhi, D. Golberg, Synthetic routes and formation mechanisms of spherical boron nitride nanoparticles, *Advanced Functional Materials*, 2008, vol. 18, no. 22, pp. 3653–3661.

[12] I. Beltagy, K. Lo, A. Cohan, SciBERT: A Pretrained Language Model for Scientific Text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3613–3618.

[13] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to Sequence Learning with Neural Networks, *arXiv*, 2014, art. no. 1409.3215.

[14] A.K. Mishra, J. Renganathan, A. Gupta, Volatility forecasting and assessing risk of financial markets using multi-transformer neural network based architecture, *Engineering Applications of Artificial Intelligence*, 2024, vol. 133, art. no. 108223.

[15] C. Carla, A.S. Uban, SciBERT meets contrastive learning: A solution for scientific hallucination detection, in: *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 336–343.

[16] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[17] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, *Neural Computation*, 1989, vol. 1, no. 4, pp. 541–551.

[18] J. Frenkel, On the transformation of light into heat in solids. I, *Physical Review*, 1931, vol. 37, no. 1, pp. 17–44.

[19] L. Pauling, The nature of the chemical bond. Application of results obtained from the quantum mechanics and from

a theory of paramagnetic susceptibility to the structure of molecules, *Journal of the American Chemical Society*, 1931, vol. 53, no. 4, pp. 1367–1400.
[20] C. Tantardini, A.R. Oganov, Thermochemical electronegativities of the elements, *Nature Communications*, 2021, vol. 12, art. no. 2087.
[21] W.G. Fateley, N.T. McDevitt, F.F. Bentley, Infrared and Raman selection rules for lattice vibrations: The correlation method, *Applied Spectroscopy*, 1971, vol. 25, no. 2, pp. 155–173.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is All you Need, *arXiv*, 2017, art. no. 1706.03762.
[23] Y. Bengio, Learning Deep Architectures for AI, *Foundations and Trends® in Machine Learning*, 2009, vol. 2, no. 1, pp. 1–127.
[24] S. Ruder, An overview of gradient descent optimization algorithms, *arXiv*, 2016, art. no. 1609.04747.
[25] X. Zhong, B. Gallagher, S. Liu, B. Kailkhura, A. Hiszpanski, T.Y.-J. Han, Explainable machine learning in materials science, *npj Computational Materials*, 2022, vol. 8, art. no. 204.

УДК 004.852

Разработка ИИ-инструмента для систематического просмотра научных статей, содержащих ИК-спектры нано- и микрочастиц нитрида бора

И.М. Соснин, А.Р. Резникова, А.В. Фролова

Лаборатория химических методов анализа элементного состава материалов, Тольяттинский государственный университет, ул. Белорусская, 14, Тольятти, 445020, Россия

Аннотация. В данной работе представлен метод создания нейронной сети и его применение для изучения данных об оптических свойствах нано- и микрочастиц гексагонального нитрида бора. В частности, метод анализирует данные о поглощении электромагнитного излучения в инфракрасной области. В работе показано, как современные алгоритмы обработки естественного языка и глубокого обучения могут быть использованы для автоматизации поиска и анализа необработанных данных. В работе мы применяем глубинные нейросетевые модели, включая сверточную нейронную сеть (CNN) для анализа инфракрасных спектров и трансформаторы (SciBERT, ChemBERTa) для анализа текстовой информации. Мультимодальное обучение, объединяющее CNN и семантический анализ текстов, было разработано для обработки гетерогенных данных.

Ключевые слова: ИК-спектр; нитрид бора; необработанные данные; CNN; ChemBERT