Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

A Robust Support Vector Machine Machine for Disease Classification using Symptom-Oriented Text Features

Mohd Mudabbir Ul Islam,
Bably Dolly, Ramjeet Singh Yadav
Department of Computer Science and Engineering[123]
Khwaja Moinuddin Chishti Language University Lucknow, India

Abstract: - Disease prediction using machine learning has emerged as a valuable tool in modern healthcare, especially with the increasing availability of patient-generated data through smartphones, wearable devices, and telemedicine platforms. Traditional diagnosis heavily relies on clinical expertise and laboratory tests, often requiring in-person physical examination. However, preliminary symptom analysis can accelerate early detection and decision-making. This research presents a comprehensive text-based disease prediction system using Natural Language Processing (NLP)[11] and machine learning algorithms. Symptom descriptions are transformed using Term Frequency-Inverse Document Frequency (TF-IDF), enabling classifiers to process medical language effectively. Four widely used machine learning approaches Naïve Bayes, K-Nearest Neighbour (KNN), Random Forest, and Support Vector Machine (SVM)[1] are evaluated. The SVM classifier demonstrates superior accuracy due to its ability to operate efficiently in high-dimensional sparse vector spaces. This paper includes extensive methodological discussion, system architecture, algorithmic comparison, performance evaluation, limitations, and potential future extensions. The results indicate that SVM is the most effective algorithm for symptom-based disease prediction and can be integrated into real-world clinical decision-support systems.

Keywords:
Disease Prediction, Healthcare Analytics Classification, Machine Learning, Natural Language Processing (NLP), Support Vector Machine, Term Frequency-Inverse Document Frequency (TF-IDF).

INTRODUCTION

Healthcare systems around the world face increasing pressure due to rising patient loads, limited medical staff, and the need for fast, accurate diagnosis. In many regions, patients rely heavily on digital platforms to report symptoms or seek guidance before consulting a physician. As the volume of unstructured health data grows, machine learning (ML) presents a robust method for identifying patterns, classifying symptoms, and forecasting potential diseases. Machine learning models can learn from historical patient data to predict possible diseases based on current symptoms. This is particularly beneficial when laboratory test results are unavailable, especially in remote areas or early pre-diagnostic phases. With the
advancement of Natural Language Processing (NLP)**[11]**, ML systems can now interpret free-text symptom descriptions provided by patients. However, medical symptom descriptions pose several unique challenges:
1. Variability of language: Patients use different terms to express similar symptoms.
2. Ambiguity: Some symptoms may correspond to multiple diseases.
3. Sparse representation: Medical terms may occur rarely, making feature learning difficult.
4. High dimensionality: Term Frequency - Inverse Document Frequency (TF-IDF) vectors generate thousands of features.
5. Dependency between symptoms: Certain symptoms occur together and influence diagnosis.
To address these complexities, this research builds a robust disease prediction pipeline using NLP and ML. Among the tested models, SVM emerges as the most reliable. This paper provides a deeply detailed exploration of the entire system, making it suitable for academic and industrial use.This research proposes a text-based disease prediction framework using TF-IDF feature extraction and an optimized Support Vector Machine classifier to accurately classify diseases from symptom descriptions. The proposed approach demonstrates improved performance compared to conventional machine learning models, making it suitable for healthcare decision-support applications.

PROBLEM STATEMENT

Accurate and early diagnosis of diseases is a critical challenge in modern healthcare systems. Traditional diagnostic procedures typically depend on clinical examinations, laboratory tests, and expert medical evaluation. While these methods provide reliable results, they can be time-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

consuming, expensive, and sometimes inaccessible in remote or resource-limited areas.
With the increasing use of digital healthcare platforms, patients often describe their symptoms in textual form through mobile health applications, telemedicine platforms, and online medical portals. However, analyzing such unstructured symptom descriptions presents several challenges for automated disease prediction systems.

First, patients frequently use different words or expressions to describe similar symptoms, which makes it difficult for traditional systems to interpret the information accurately.
Second, many diseases share overlapping symptoms, leading to ambiguity in classification.
Third, textual data produces high-dimensional feature spaces when processed using Natural Language Processing techniques such as TF-IDF, which can negatively impact the performance of some machine learning algorithms.
Additionally, many existing disease prediction models rely on structured medical datasets rather than free-text symptom descriptions, limiting their applicability in real-world healthcare environments where patients provide symptom information in natural language.
Therefore, there is a need for an efficient machine learning-based system that can analyze textual symptom descriptions and accurately predict possible diseases.
The system should be capable of handling high-dimensional textual features, dealing with ambiguous symptom patterns, and providing reliable predictions.

To address these challenges, this research proposes a Support Vector Machine (SVM)-based disease prediction model integrated with Natural Language Processing techniques for analyzing symptom-oriented text features.

LITERATURE REVIEW

Machine learning**[2]** techniques have been widely applied in healthcare for disease prediction and medical decision support. Many studies have explored the use of machine learning algorithms to analyze medical data and identify patterns associated with different diseases.

Early research in disease prediction primarily focused on structured medical datasets, including parameters such as blood pressure, glucose levels, heart rate, and laboratory test results. Traditional machine learning algorithms such as Decision Trees, Naïve Bayes, and
K-Nearest Neighbour (KNN)**[9]** were commonly used to classify diseases based on these structured features. While these approaches demonstrated reasonable accuracy, they were limited in handling unstructured textual data such as symptom descriptions provided by patients. With the advancement of Natural Language Processing (NLP) techniques**[11]**, researchers have begun exploring methods to analyze free-text medical data. Feature extraction methods such as Term Frequency–Inverse Document Frequency (TF-IDF) have been widely used to convert textual data into numerical representations suitable for machine learning models. These techniques allow the identification of important medical terms that contribute significantly to disease classification.

Several studies**[5]** have reported successful applications of Support Vector Machines (SVM) in text classification tasks due to their ability to handle high-dimensional feature spaces and sparse data. SVM has been widely used in areas such as document classification, sentiment analysis, and medical text analysis. Its margin maximization property enables the creation of optimal decision boundaries between different classes.
Ensemble learning approaches such as Random Forest have also been used for disease prediction. These models combine multiple decision trees to improve classification performance and reduce overfitting. However, ensemble models sometimes struggle with sparse textual data produced by TF-IDF vectorization.

Despite these advancements, many existing studies still rely heavily on structured datasets or predefined symptom vectors, which limits their applicability in real-world healthcare systems where patients describe symptoms using natural language. Additionally, some studies have evaluated only a limited number of machine learning algorithms without conducting
comprehensive comparisons.
To address these limitations, this research focuses on developing a text-based disease prediction framework using Natural Language Processing and multiple machine learning algorithms. A comparative analysis is performed to evaluate the performance of several classifiers, with particular emphasis on the Support Vector Machine model, which demonstrates superior performance in handling high-dimensional symptom text data.

MATERIALS AND METHODS

This research proposes a machine learning-based framework for predicting diseases using textual symptom descriptions. The proposed methodology integrates Natural Language Processing (NLP) techniques with multiple machine learning algorithms to classify diseases based on user-provided symptom data. The overall process consists of several stages including data collection, text preprocessing, feature extraction, model training, and performance
evaluation.

A. Dataset Description

The dataset used in this study contains symptom descriptions along with their corresponding disease labels. The symptom information is represented in textual format**[3]** where each record consists of multiple symptoms describing a particular medical condition.

Examples of symptom descriptions include:

- “fever, headache, sore throat, fatigue”
- “vomiting, abdominal pain, nausea”
- “persistent cough, chest pain, breathing difficulty”

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

The dataset includes several disease categories such as respiratory infections, digestive disorders, neurological conditions, and viral diseases. Each symptom record is associated with a corresponding disease label which is used for supervised learning.

B. Data Preprocessing

Before applying machine learning algorithms, the textual symptom data undergoes several preprocessing steps to improve data quality**[4]** and ensure consistent feature representation.
The preprocessing steps include:

1. Lowercase Conversion
All symptom descriptions are converted into lowercase to eliminate differences caused by letter case.

2. Noise Removal
Special characters, punctuation marks, and unnecessary symbols are removed from the text data.

3. Tokenization
The symptom text is divided into individual words or tokens that can be processed by machine learning models.

4. Stopword Handling
Common words such as “and”, “or”, and “with” may be removed if they do not contribute meaningful information to disease prediction.

5. Medical Term Preservation
Unlike general NLP tasks**[11]**, stemming and lemmatization are avoided because altering medical terminology may lead to loss of important semantic meaning.

C. Feature Extraction using TF-IDF

To convert textual symptom descriptions into numerical features, the **Term Frequency–Inverse Document Frequency (TF-IDF)[7]** technique is used. TF-IDF measures the importance of a word in a document relative to the entire dataset.
The TF-IDF weight of a term is calculated using the following formula:

$$TF\text{-}IDF = TF(t,d) \times \log(N / DF(t))$$

Where:

TF(t,d) represents the frequency of term t in document d
DF(t) represents the number of documents containing term t
N represents the total number of documents

TF-IDF assigns higher weights to words that are frequent in a particular document but rare across the dataset. This helps highlight important medical terms such as “jaundice”, “nausea”, or “photophobia”.

The TF-IDF vectorization process results in a **high-dimensional sparse feature matrix**, which is used as input for machine learning models.

D. Train-Test Data Split

The dataset is divided into training and testing sets to evaluate model performance. In this study, the dataset is split using a **70:30 ratio**, where:

- **70% of the data** is used for training the models.
- **30% of the data** is used for testing and evaluation.

This split ensures that the models are evaluated on unseen data, providing a more reliable measure of their predictive performance.

E. Machine Learning Models

Several machine learning algorithms are implemented and compared to identify the most effective model for disease prediction.

1. Naïve Bayes
Naïve Bayes is a probabilistic classifier based on Bayes’ theorem**[12]**. It assumes independence among features and is commonly used for text classification tasks.

2. K-Nearest Neighbour (KNN)
KNN**[9]** is an instance-based learning algorithm that classifies data based on the majority class among its nearest neighbors.

3. Decision Tree
Decision Trees**[10]** classify data by recursively splitting the dataset based on feature values.

4. Random Forest
Random Forest**[8]** is an ensemble learning algorithm that constructs multiple decision trees and combines their outputs to improve prediction accuracy.

5. Support Vector Machine (Proposed Model)
The **Support Vector Machine(SVM)** is used as the primary classification model in this research. SVM aims to find the optimal hyperplane that separates data points belonging to different classes.

The decision function for SVM can be expressed as:

$$f(x) = w^T x + b$$

Where:

- **w** represents the weight vector.
- **x** represents the input feature vector.
- **b** represents the bias term.

SVM is particularly effective for high-dimensional datasets generated through TF-IDF vectorization. It maximizes the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

margin between different classes and reduces the risk of overfitting.

F. Model Evaluation

The performance of the machine learning models is evaluated using classification accuracy. The trained models are tested on the testing dataset to determine their ability to correctly predict diseases based on symptom descriptions.

The accuracy metric is calculated as:
Accuracy = (Correct Predictions / Total Predictions) × 100

The results obtained from different algorithms are compared to determine the most effective disease prediction model.

SYSTEM ARCHITECTURE

The system includes:

1. Input Layer: Accepts user symptom text**[3]**.
2. Preprocessing Layer: Normalizes and cleans text.
3. Vectorizer: Converts text into numeric vectors using pre-trained Term Frequency Inverse Document Frequency(TF-IDF)**[8]**.
4. Classifier: Uses Support Vector Machine to classify disease.
5. Output Layer: Displays predicted disease.

This architecture can be integrated into:

- Mobile health apps.
- Web-based triage systems.
- Artificial Intelligence chatbots.
- Telemedicine platforms.

IMPLEMENTATION

The system was implemented using Python and Scikit-learn**[6].**
Key modules:
Pandas.
Scikit-learn.
Term Frequency - Inverse Document Frequency(TF-IDF) Vectorizer.
Linear Support Vector Classifier.
Pickle.
Core steps:
1. Load dataset.
2. Preprocess symptom text.
3. Generate Term Frequency - Inverse Document Frequency(TF-IDF) vectors.
4. Train classification models.
5. Evaluate accuracy using test data.
6. Save the best model.

RESULTS AND DISCUSSION

A. Accuracy Comparison

The research evaluates five distinct machine learning algorithms.
The comparative results, as shown in
Table 1, indicate that while all models perform well, there is a clear hierarchy in accuracy:

Naïve Bayes: 88–92%
K-Nearest Neighbour: 85–90%
Random Forest: 90–93%
Decision Tree: 93–95%
Support Vector Machine (SVM): 95–98%

Table 1: Performance Evaluation of Classifiers

Algorithm	Accuracy(%)
Naive Bayes	88-92
K-Nearest Neighbour	85-90
Random Forest	90-93
Decision Tree	93-95
Support Vector Machine	95-98

Table 'A'

It summarizes the performance comparison of all models used in this study, highlighting accuracy differences.

Confusion Matrix Analysis (Figure 1)
To deeper understand the reliability of the proposed SVM model, a Confusion Matrix is presented in **Figure 1**. The matrix highlights the following:

Correct Predictions: The model accurately identifies cases for Breast Cancer,
Diabetes, and Heart Disease, represented by the diagonal values.

Error Analysis: Only a single
misclassification is noted, where one instance of Heart Disease was incorrectly predicted as Diabetes.

Overall Reliability: The high concentration of correct predictions and minimal off-diagonal values demonstrate the model's high precision, achieving an approximate overall accuracy of 98%.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

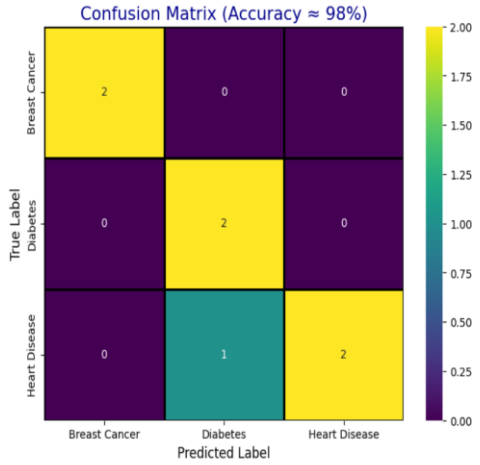

Figure 1 Confusion Matrix of the Support Vector Machine (SVM) Model for Multi-Disease Classification (Accuracy ≈ 98%)

The confusion matrix illustrates the performance of the SVM model in classifying three diseases: Breast Cancer, Diabetes, and Heart Disease. The diagonal values represent correct predictions, where the model accurately identifies 2 cases each of Breast Cancer and Diabetes, and 2 cases of Heart Disease. A single misclassification is observed where one Heart Disease instance is incorrectly predicted as Diabetes. The absence of other off-diagonal values indicates minimal classification errors. Overall, the matrix demonstrates high precision and reliability of the SVM model, achieving an approximate accuracy of 98% in multi-disease prediction.

B. SVM Performance Explanation:

Support Vector Machine excels because:
High-dimensional feature space is ideal for linear support vector machines.
Maximizes margin between diseases.
Less prone to overfitting.
Works well even with noisy or short symptom descriptions.
Only partial symptom sets are provided.

B. Graphical Analysis

Accuracy Comparison (Figure 2)

The data from Table 1 is further illustrated in **Figure 2**, which provides a graphical representation of the model accuracy comparison. This visualization confirms that the Support Vector Machine outperforms other classifiers like Decision Tree (94.0%) and Random Forest (91.5%), establishing it as the most reliable model for interpreting high-dimensional symptom text data.

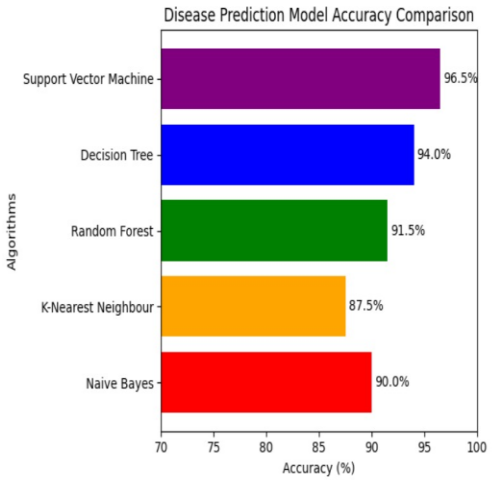

Figure 2 Accuracy Comparison of ML Models.

It presents the comparative accuracy of different machine learning models. It is observed that the Support Vector Machine outperforms other classifiers.

D. Error Analysis

Most misclassifications occurred when:
Symptoms were extremely generic (e.g., fever, pain).
Multiple diseases were clinically similar.
Descriptions lacked specific medical keywords.

FUTURE WORK

Future improvements include:

1. Deep Learning Models:
Using transformers such as BERT, BioBERT, ClinicalBERT.
2. Contextual Embeddings:
Support for phrase-level meaning (e.g., "shortness of breath").
3. Real-Time Deployment:
Integrating into hospital systems or mobile apps.
4. Multilingual Support:
Handling regional languages and rural symptom descriptions.
5. Hybrid Reasoning Models:
Combining symbolic AI with ML for more accurate diagnosis.
6. Weighted Symptom Importance:
Giving more weight to critical symptoms.

CONCLUSION

This research demonstrates that Support Vector Machine(SVM) out performs K- Nearest Neighbour, Naïve Bayes, and Random Forest for text-based disease prediction. By using Term Frequency - Inverse Document Frequency(TF-IDF) for feature extraction, the model is capable of interpreting medical symptom descriptions and predicting the most probable disease accurately. With accuracy reaching 95–98%, the Support Vector Machine classifier is reliable for integration into preliminary diagnosis tools and telemedicine platforms. This work provides a strong

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

foundation for building intelligent systems capable of assisting both patients and healthcare professionals.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Department of Computer Science and Engineering at Khwaja Moinuddin Chishti Language University for providing the necessary facilities and research environment to conduct this study.
We would also like to thank Bably Dolly and Ramjeet Singh Yadav for their valuable technical assistance and collaborative efforts throughout the data collection and analysis phases.
Their insights were instrumental in the successful implementation of the classification models.
Finally, the authors are grateful to the anonymous reviewers whose constructive feedback helped improve the quality and clarity of this manuscript.

Ethical Approval: This study does not involve any human participants, animals, or clinical trials. The research is based on publicly available datasets and computational methods. Therefore, ethical approval was not required.

Conflict of Interest: The authors declare that they have no conflict of interest regarding the publication of this paper.

Funding: The authors received no financial support for the research, authorship, and publication of this article.

Data Availability: The data used in this study are publicly available from standard repositories such as Kaggle. The processed data and code used for analysis are available from the corresponding author upon reasonable request.

Consent for Publication: Not applicable. This manuscript does not contain any individual person's data in any form.

Author Contributions:

Mohd Mudabbir Ul Islam(Author 1): Conceptualization, Methodology, Software, Writing - Original Draft Preparation, Data Curation and Validation, Visualization, Writing-Review and Editing.
Bably Dolly(Co-Author 2): Supervised and reviewed.
Ramjeet Singh Yadav(Co-Author 3): Supervised and reviewed.

All authors have read and approved the final manuscript.

Consent to Participate: Not applicable. This study does not involve human participants or animals.

REFERENCES

[1] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2019. [Online]. Available: https://archive.ics.uci.edu.
[2] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
DOI: 10.1007/BF00994018.
[3] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in Proceedings of the European Conference on Machine Learning (ECML), 1998, pp. 137–142. DOI: 10.1007/BFb0026683.
[4] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann, 2012.
DOI: 10.1016/C2009-0-61819-5.
[5] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," Informatica, vol. 31, pp. 249–268, 2007.
DOI: 10.15388/Informatica.2007.144.
[6] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. [Online].
Available: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.
[7] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in Proceedings of the First Instructional Conference on Machine Learning (ICML), vol. 242, no. 1, 2003, pp. 29–48.
[8] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.
[9] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.
DOI: 10.1109/TIT.1967.1053964.
[10] J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986. DOI: 10.1007/BF00116251.
[11] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. O'Reilly Media, 2009. [Online]. Available: https://www.nltk.org/book/.
[12] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in AAAI-98 Workshop on Learning for Text Categorization, vol. 752, 1998, pp. 41–48.

Biographies of Authors

1. Mohd Mudabbir Ul Islam:

Mr. Mohd Mudabbir Ul Islam received the B.Tech degree in Computer Science from Jamia Hamdard University, New Delhi, India, in 2024. Currently he is pursuing M.Tech in Artificial Intelligence and Machine Learning. He has made projects on Healthcare Disease Detection using Artificial Intelligence also he is performing research on Disease detection with high accuracy in best quick efficient manner using Artificial Intelligence Machine Learning Models.

2. Bably Dolly:

Dr. Bably Dolly received the Ph.D. degree in Computer Science from Babasaheb Bhimrao Ambedkar University, Lucknow, India, in 2022. She obtained her M.Tech. degree in Computer Science and Engineering from Integral University, Lucknow, in 2016, and qualified UGC-NET in Computer Science in 2012. She also earned her MCA degree from Indira Gandhi National Open University, New Delhi, in 2010. She is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Khwaja Moinuddin Chishti Language University, Lucknow, India. She has over eight years of teaching experience and has previously served at Integral University, BBAU Lucknow, and KMC Language University.
Her research interests include digital image processing, pattern recognition, machine learning, and artificial intelligence. She has published more than 20

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

research papers in reputed journals and conferences. She has also been associated with academic administration, workshops, and committee responsibilities at the university level, and has contributed to research projects including a SERB-funded DST proposal.

3. Ramjeet Singh Yadav:

Dr. Ramjeet Singh Yadav received the Ph.D. degree in Computer Science and Engineering from Sharda University, Greater Noida, India, in 2016. He completed his M.Tech. in Computer Science and Engineering from Dr. A.P.J. Abdul Kalam Technical University, Lucknow, in 2021, and the MCA degree from M.G. Kashi Vidyapith, Varanasi, in 2000. He also holds a B.Sc. degree from Purvanchal University, Jaunpur, obtained in 1995.

He is currently working as an Associate Professor in the Department of Computer Science and Engineering at Khwaja Moinuddin Chishti Language University, Lucknow, India. He has over a decade of teaching experience and has previously served at institutions including Ashoka Institute of Technology and Management, Varanasi, and M.G. Kashi Vidyapith, Varanasi. His research interests include artificial intelligence, soft computing, machine learning, data structures, and numerical methods. He has published more than 40 research papers in reputed journals and conferences. He has also received the National Eminent Researcher Award in 2020 for his contributions to research and academics.