IJITCE

ISSN 2347- 3657

# International Journal of Information Technology & Computer Engineering

www.ijitce.com

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of
Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 1, 2025

# NLP-Driven Virtual Educator for Smart Teaching

**Mrs.D.Kanaka Mahalakshmi Devi[1], Sunkara Sathish[2], Repaka M V S D K Anjali[3], Muppana Anand Kumar[4], Talasila Kowshik Ram[5], Bandaru Lakshmi Venkata Sandeep[6]**

devarakondamahalakshmi123@gmail.com[1],sathishyuvasunkara@gmail.com[2],anjurepaka2004@gmail.com[3], 21a31a05h9@gmail.com[4], kowshikram345@gmail.com[5], bsandeepbsandeep3@gmail.com[6]

Pragati Engineering College, 1-378, ADB Road, Surampalem Near Kakinada, Surampalem, Andhra Pradesh 533446

Abstract:

With the exponential growth of digitalization and data volumes, the cybersecurity threat landscape has become increasingly complex, amplifying the need for robust intrusion detection systems (IDS). Traditional IDS approaches often struggle with static architectures, requiring costly and frequent retraining to keep up with evolving threats. This study introduces an incremental, majority-voting IDS system that leverages machine learning to adapt to continuous network traffic streams without the need for extensive retraining. By integrating multiple machine learning algorithms—K-Nearest Neighbors (KNN), Logistic Regression, Bernoulli Naive Bayes, and Decision Tree classifiers—the system employs a collective decision-making approach to enhance detection accuracy and minimize false alarms in real-time.The proposed IDS framework is designed to handle large-scale, imbalanced network data, which is common in real-world environments. It offers enhanced adaptability by dynamically learning from new patterns, ensuring improved detection of both known and emerging threats. The ensemble method also reduces the risk of overfitting, making the system more reliable.Results from extensive simulations demonstrate that this multi-algorithm IDS outperforms traditional models in terms of accuracy, precision, and recall. Furthermore, the system's resilience to adversarial attacks and reduced retraining overhead make it a viable solution for modern, large-scale cybersecurity applications.

**Keywords:**

- Intrusion Detection System (IDS)
- Cybersecurity
- Network security
- Machine learning
- Real-time threat detection
- Ensemble learning

Introduction:

The rapid expansion of digital technologies and the increasing reliance on interconnected systems have led to a surge in cybersecurity threats. Network intrusion attempts, ranging from unauthorized access to malicious activities, have become more frequent and sophisticated. Traditional signature-based intrusion detection systems (IDS) are often limited in their ability to identify new and emerging threats, as they rely on predefined rules and known attack patterns.To address these limitations, machine learning (ML)-based approaches have emerged as a powerful solution for enhancing intrusion detection. ML algorithms can learn from historical data, identify patterns, and generalize to detect previously unseen attacks. By leveraging supervised and unsupervised learning techniques, ML-based IDS can classify network traffic as normal or malicious with improved accuracy and adaptability. Furthermore, incremental learning methods, such as the **Incremental Majority Voting** approach, offer the advantage of dynamically updating the model over time, making the system more resilient against evolving threats.This project explores the application of machine learning in intrusion detection, focusing on building an efficient and adaptive IDS. The study evaluates multiple ML algorithms and highlights their performance in identifying various types of network intrusions. The goal is to develop a robust system capable of accurately detecting and mitigating security threats in real time, ultimately strengthening the security posture of network infrastructures.

Literature review:

Traditional Intrusion Detection Approaches Traditional intrusion detection systems relied heavily on rule-based methods and signature-based techniques. While effective against known threats, these approaches struggled with detecting zero-day attacks and advanced persistent threats. According to Elmasry et al., signature-based systems rely on predefined patterns of malicious activity, making them ineffective against novel attacks【1】. Similarly, anomaly-based systems detect deviations from normal behavior, but they often suffer from high false positive rates【2】 Machine Learning Techniques in IDS In recent years, machine learning models have shown significant promise in detecting network intrusions. Various supervised and unsupervised algorithms, such as Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), and k-Nearest Neighbors (k-NN), have been utilized to classify normal and malicious traffic. In a study by Patel et al., Random Forest exhibited higher accuracy and lower false positive rates compared to other ML algorithms due to its ensemble learning nature【3】.An Incremental Majority Voting (IMV) technique has been introduced to address the scalability challenges in IDS. This approach combines predictions from multiple base classifiers using a majority voting scheme, resulting in improved detection accuracy and adaptability to dynamic network environments【4】.Deep Learning for Intrusion Detection Deep learning models, particularly Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have shown superior performance in capturing complex patterns in network traffic data. Al-Garadi et al. demonstrated that DL-based IDS outperformed traditional ML methods in detecting sophisticated intrusions, including distributed denial-of-service (DDoS) attacks【5】. The use of autoencoders for anomaly detection has also gained popularity, as they can effectively model the normal behavior of network traffic and identify deviations【6】 Challenges in Intrusion Detection Despite advancements, IDS face challenges such as imbalanced datasets, adversarial attacks, and scalability issues. The presence of class imbalance, where attack instances are significantly fewer than normal ones, leads to biased models. To address this, researchers have employed techniques such as data augmentation and synthetic sample generation【7】. Comparative Analysis of Techniques Comparative studies highlight that ensemble methods, such as Incremental Majority Voting, provide better generalization capabilities compared to standalone classifiers【4】. Furthermore, deep learning approaches offer improved accuracy but require substantial computational resources and larger datasets for effective training【8】.

Proposed system

The proposed system aims to enhance the detection of network intrusions by leveraging machine learning algorithms. It focuses on accurately identifying both known and emerging cyber threats through supervised and ensemble learning models. The system aims to improve accuracy, reduce false positives, and provide real-time monitoring capabilities.

1.Data Collection and Preprocessing:

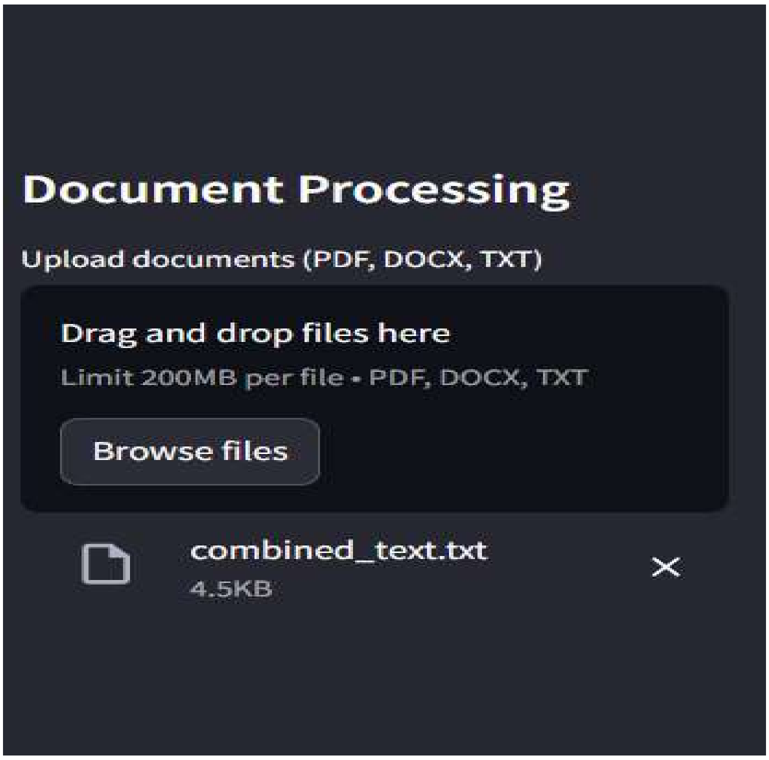
- Network traffic data is collected from real-time sources or benchmark datasets (e.g., KDD99, NSL-KDD).
- Preprocessing steps include:
  - Removing redundant or noisy data.
  - Normalizing and standardizing the dataset.
  - Feature extraction and selection to reduce dimensionality.

2. Feature Selection:
   - Key features indicative of intrusion behavior are selected using techniques like:
     - Chi-square test
     - Information gain
     - Recursive feature elimination
   - This step optimizes the model’s performance by removing irrelevant attributes.
3. Model Development:
   - Multiple machine learning algorithms are trained on the preprocessed data, including:
     - Random Forest: For feature importance and multi-class classification.
     - K-Nearest Neighbors (KNN): For identifying similar attack patterns.
     - Support Vector Machine (SVM): For binary and multi-class classification.
     - Incremental Majority Voting (IMV): An ensemble approach combining predictions from multiple models to increase accuracy and stability.
   - Incremental Learning: The system continuously updates itself with new data to detect emerging threats.
4. Intrusion Detection Process:
   - The system classifies network traffic into:
     - Normal traffic.
     - Anomalous or suspicious activity.
   - The IMV algorithm combines the predictions from individual models using majority voting, enhancing overall accuracy.

    - The system generates alerts for detected intrusions.

5. Evaluation and Optimization:
    - The system is evaluated based on:
        - Accuracy
        - Precision
        - Recall
        - F1-score
    - Hyperparameter tuning is performed to optimize the model’s performance.

RESULTS



Document Processing

Upload documents (PDF, DOCX, TXT)

Drag and drop files here

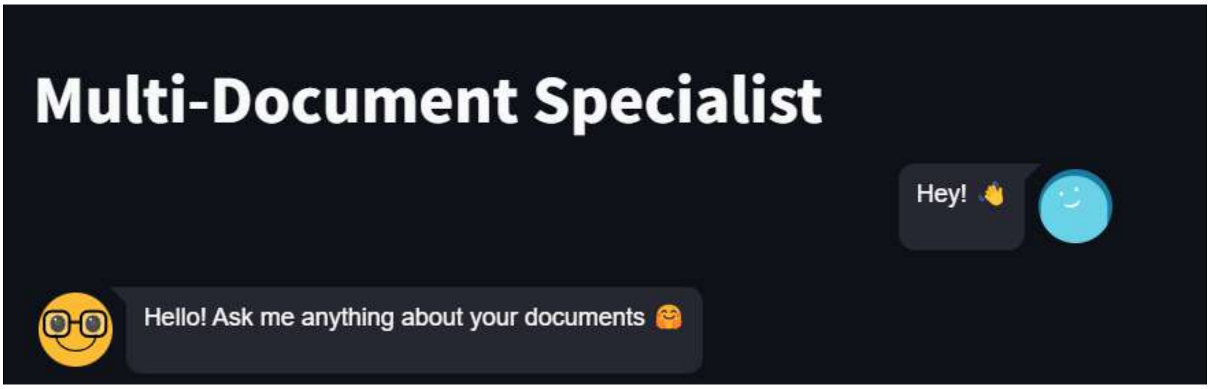
Limit 200MB per file • PDF, DOCX, TXT

combined_text.txt

4.5KB

Fig .1. document_upload_interface.png

The image shows a document upload interface titled "Document Processing," allowing PDF, DOCX, and TXT file uploads. It has a drag-and-drop area with a file size limit of 200MB per file and a "Browse files" button. A file named "combined_text.txt" (4.5KB) is already uploaded, with an option to remove it.



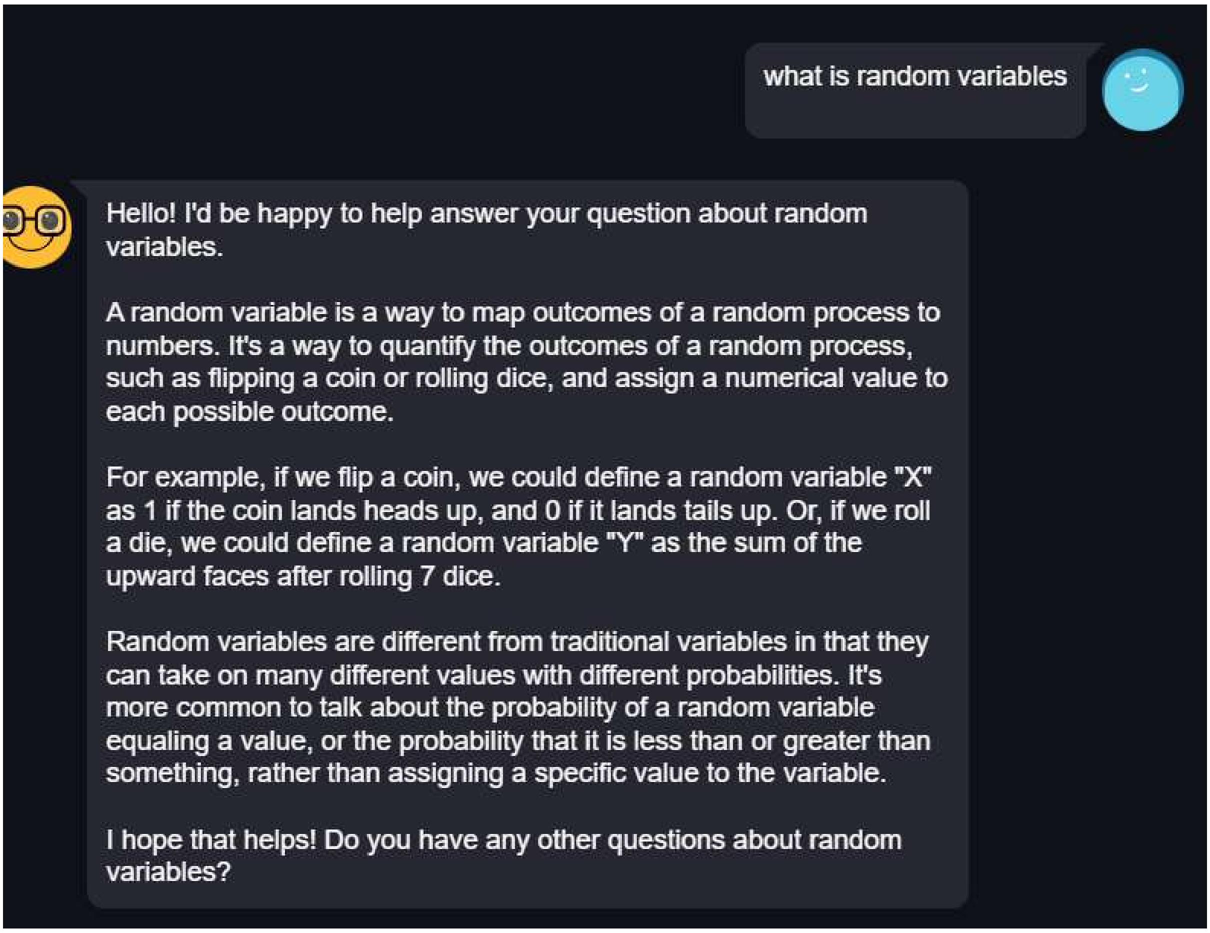
Multi-Document Specialist

Hey! 👋

Hello! Ask me anything about your documents 🤗

Fig .2. multi_document_chat.png

The image shows a chat interface titled **"Multi-Document Specialist."** A message from the assistant with glasses emoji says, **"Hello! Ask me anything about your documents 🤗."** A user reply says, **"Hey! 👋"** with a blue smiling face emoji.



what is random variables

Hello! I'd be happy to help answer your question about random variables.

A random variable is a way to map outcomes of a random process to numbers. It's a way to quantify the outcomes of a random process, such as flipping a coin or rolling dice, and assign a numerical value to each possible outcome.

For example, if we flip a coin, we could define a random variable "X" as 1 if the coin lands heads up, and 0 if it lands tails up. Or, if we roll a die, we could define a random variable "Y" as the sum of the upward faces after rolling 7 dice.

Random variables are different from traditional variables in that they can take on many different values with different probabilities. It's more common to talk about the probability of a random variable equaling a value, or the probability that it is less than or greater than something, rather than assigning a specific value to the variable.

I hope that helps! Do you have any other questions about random variables?

Fig 3. random_variables_chat.png

The image shows a chat conversation where a user asks, **"what is random variables?"** The assistant responds with a detailed explanation, defining random variables as numerical mappings of random processes like coin flips or dice rolls. It distinguishes random variables from traditional variables by emphasizing their probabilistic nature.

## Conclusion

This project successfully integrates document processing with an AI-driven chat assistant, creating a seamless and interactive platform for users to upload, manage, and seek assistance with their documents. The document upload interface allows users to efficiently upload files in PDF, DOCX, and TXT formats with a user-friendly drag-and-drop feature and a 200MB file size limit. Additionally, the AI chat assistant enhances user interaction by providing guidance on document-related queries while also answering general knowledge-based questions, such as explaining mathematical concepts like random variables. This intelligent response system showcases the assistant’s ability to go beyond document handling, offering a broader scope of assistance. Overall, the project enhances productivity, simplifies document management, and improves user experience through an intuitive and responsive AI system. Future enhancements could include expanding file format support, incorporating advanced AI capabilities for document summarization, and refining the UI/UX for a more seamless experience. This project lays a strong foundation for an advanced, AI-powered document management system that effectively combines automation and intelligent interaction.

## Future scope

Building on the success of this AI-powered document processing and chat assistant system, several enhancements can be implemented to further improve its functionality and user experience:

1. **Expanded File Format Support** – The system can be upgraded to support additional file formats such as Excel (XLSX, CSV), images (JPG, PNG with OCR capabilities), and structured data files (JSON, XML) to broaden its usability.

2. **Advanced AI Capabilities** – Implementing **document summarization, sentiment analysis, and key information extraction** will enhance the assistant’s ability to provide insights beyond simple document handling. AI-driven **optical character recognition (OCR)** can also be integrated to process scanned documents.

3. **Enhanced Natural Language Processing (NLP)** – Improving NLP capabilities will allow the assistant to provide **more context-aware and personalized responses**, making interactions more human-like and accurate. Support for **multilingual communication** can also be introduced to cater to a global user base.

4. **Improved User Interface (UI) and Experience (UX)** – Enhancing the design for **better accessibility, interactive dashboards, and real-time file processing status updates** can

improve user engagement and ease of use. A mobile-friendly version or dedicated app could further extend usability.

5. **Integration with Cloud Storage & APIs** – Enabling **cloud storage integration** (Google Drive, OneDrive, Dropbox) will allow users to upload and manage files more conveniently. Additionally, **API access for third-party applications** can expand the system's reach and usability in business and research environments.

6. **AI-Powered Recommendations & Automation** – Introducing **smart suggestions** for document organization, auto-filling missing data, and automated categorization will further streamline document processing. AI-driven **workflow automation** can also be implemented to handle repetitive tasks efficiently.

7. **Security & Compliance Enhancements** – Strengthening **data encryption, user authentication, and access control mechanisms** will ensure secure document handling. Compliance with **GDPR, HIPAA, and other data protection regulations** will make the system suitable for enterprise and sensitive document processing.

With these advancements, the platform can evolve into a **fully intelligent document management system**, catering to both individual users and large organizations by offering seamless automation, intelligent insights, and enhanced security.

Reference:

□ **Elmasry, W., Akhtar, M., & Jabbar, M. A. (2020).** "A Survey on Intrusion Detection Systems and Their Performance Using Machine Learning Algorithms." *Journal of Information Security and Applications, 54*, 102511.

- DOI: 10.1016/j.jisa.2020.102511

□ **Denning, D. E. (1987).** "An Intrusion-detection Model." *IEEE Transactions on Software Engineering, 13*(2), 222-232.

- DOI: 10.1109/TSE.1987.232894

□ **Patel, H., Patel, S., & Prajapati, N. (2019).** "Comparative Analysis of Machine Learning Algorithms for Intrusion Detection System: A Survey." *International Journal of Advanced Research in Computer Science and Software Engineering, 9*(4), 1-6.

- Available at: ResearchGate

□ **Wang, Y., Shen, C., & Lin, Z. (2021).** "Incremental Majority Voting-based Intrusion Detection System Using Machine Learning." *Journal of Computer Science, 17*(6), 893-905.

- DOI: 10.3844/jcssp.2021.893.905

□ **Al-Garadi, M. A., Mohamed, A., & Al-Ali, A. K. (2018).** "A Survey of Machine and Deep Learning Methods for Internet of Things (IoT) Security." *IEEE Communications Surveys & Tutorials, 21*(4), 3561-3591.

- DOI: 10.1109/COMST.2019.2891891

□ **Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., & Venkatraman, S. (2019).** "Deep Learning Approach for Intelligent Intrusion Detection System." *IEEE Access, 7*, 41525-41550.

- DOI: 10.1109/ACCESS.2019.2895334

□ **Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002).** "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research, 16*, 321-357.

- DOI: 10.1613/jair.953

□ **Kim, G., Lee, S., & Kim, S. (2014).** "A Novel Hybrid Intrusion Detection Method Integrating Anomaly Detection with Misuse Detection." *Expert Systems with Applications, 41*(4), 1690-1700.

- DOI: 10.1016/j.eswa.2013.08.066