Research Trend Analysis and Future Prediction using Deep Learning

---

Team Members:
Muhammad Talha Tahir (mtalhatahir2001@gmail.com) 2024MSCS227
Haseeb Ahmed (haaseebahmed319@gmail.com) 2024MSCS221
Muhammad Usama (utari7302@gmail.com) 2024MSCS222

Subject: Deep Learning (MSCS)

Project Title: Analysis of Research Trends from Academic Papers and Prediction of Emerging Research Areas

---

1. Introduction

In this project, we aim to extract, analyze, and predict research trends across various Computer Science domains over the past several decades. We collected a large dataset of research papers, analyzed the data for past trends and citation impact, clustered research topics, and finally built a deep learning model to predict emerging research areas.

The complete workflow covered:

- Web scraping for data collection
- Data cleaning and standardization
- Exploratory data analysis (EDA)
- Research topic clustering
- Citation impact analysis
- Future trend prediction using LSTM (deep learning)

---

2. Data Collection

For data collection, we developed a custom **Scrapy** spiders (Python framework) to automate extraction of metadata from research journal websites, primarily focusing on the Elsevier and Springer platform.

Challenges:

- **Bot Detection:** Elsevier actively detects and blocks non-human traffic. We faced captchas and IP blocks frequently.
- **Solution:** We integrated **ScraperAPI** (a proxy API service) which provided rotating proxies and user-agent spoofing. Using free credits from ScraperAPI, we were able to significantly bypass blocking mechanisms.

Dataset Details:

- **Total papers collected:** 4971
- **Coverage years:** 1984 – 2025
- **Fields extracted:**
    - Title
    - Abstract
    - Year
    - Citation count
    - Author(s) with country
    - Keywords

Note: A few journals were skipped due to ScraperAPI usage limits.

---

3. Data Cleaning

Data cleaning was performed using simple Python scripts.

Steps:

- Removal of invalid Unicode characters from titles and abstracts.
- Dropping papers with completely missing abstracts or titles.
- Standardization of year format to ensure consistency.

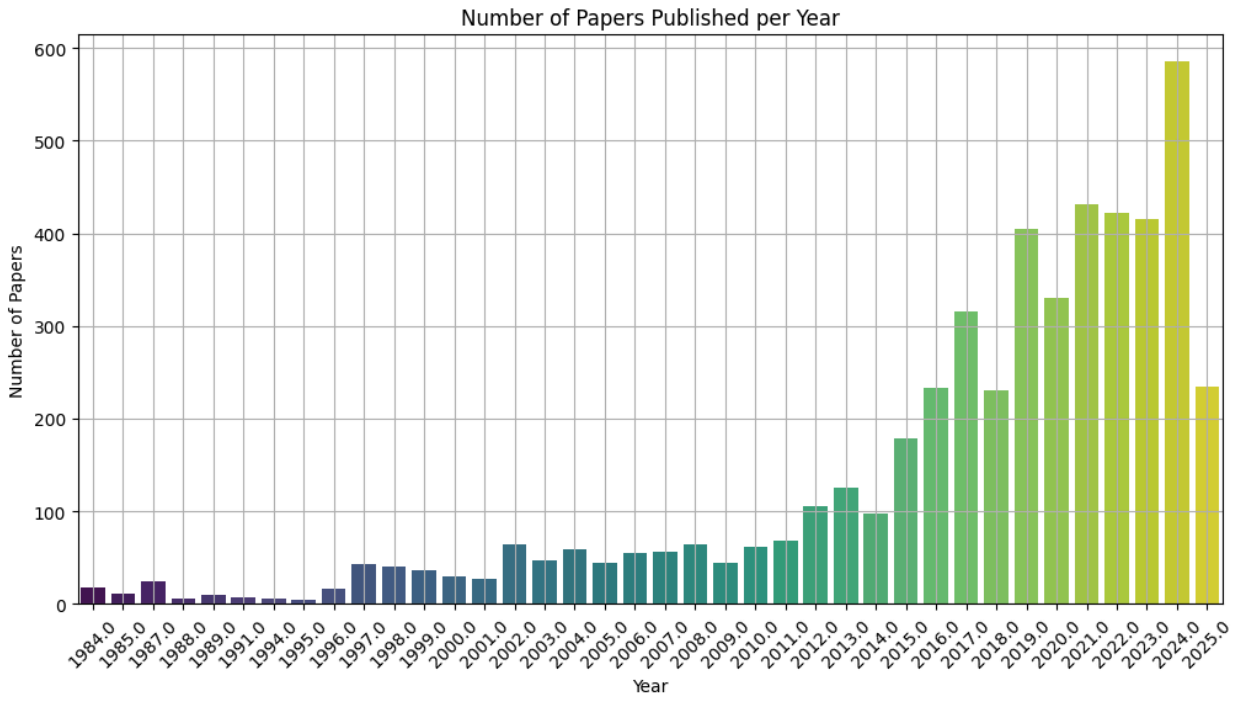
Overall, missing data was **not a major issue**; only a few records were discarded.

---

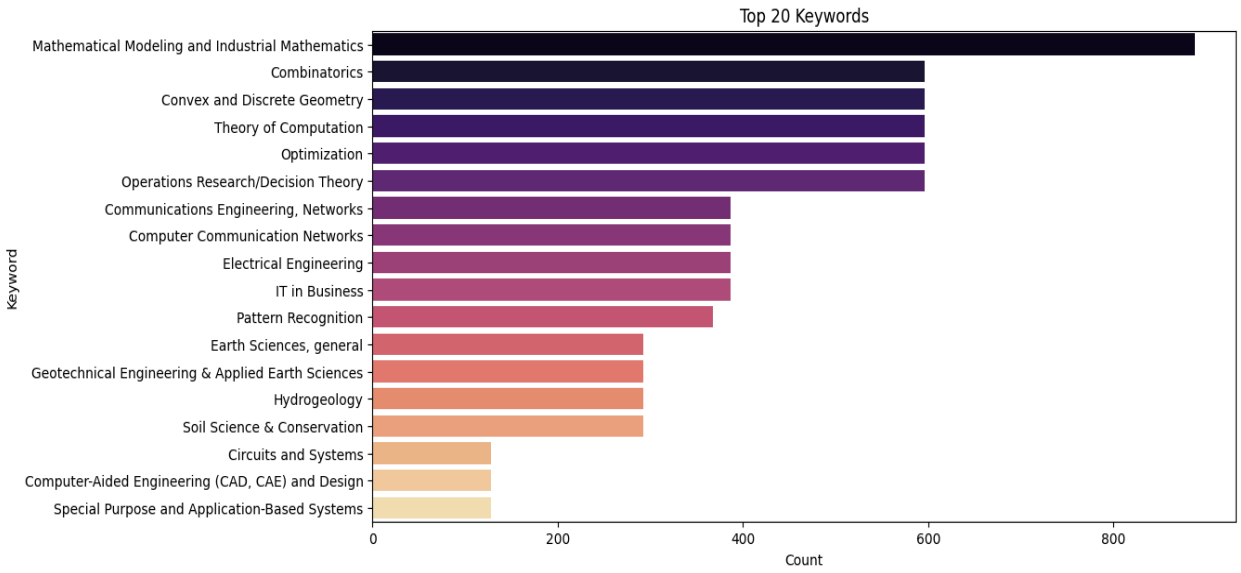
4. Exploratory Data Analysis (EDA)

Key Insights Explored:

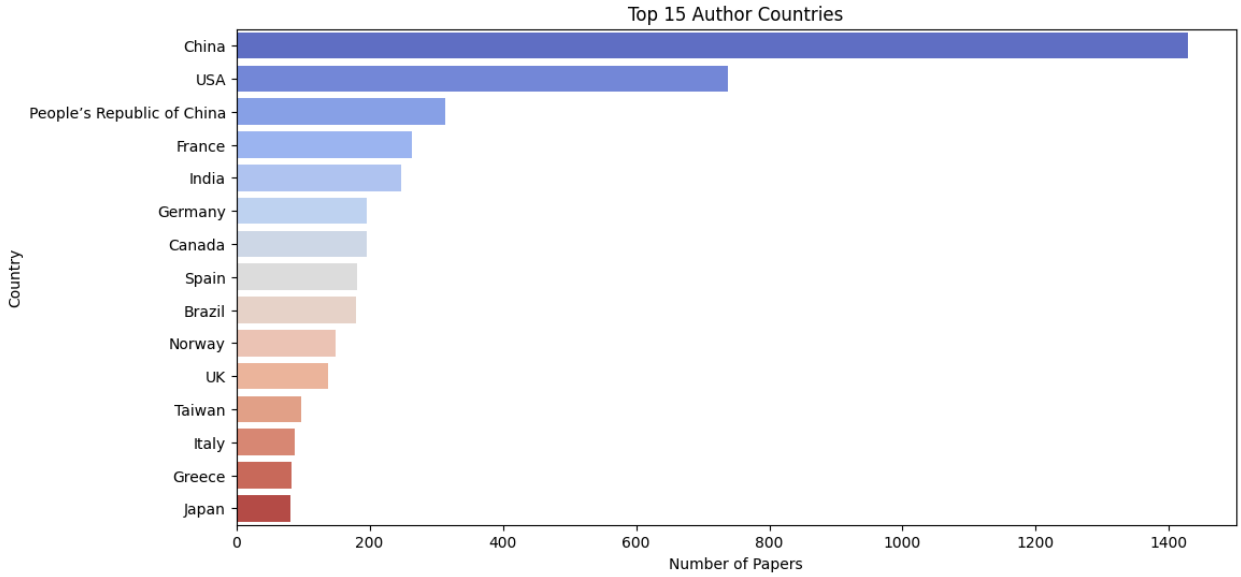
1. **Number of Papers Published per Year**



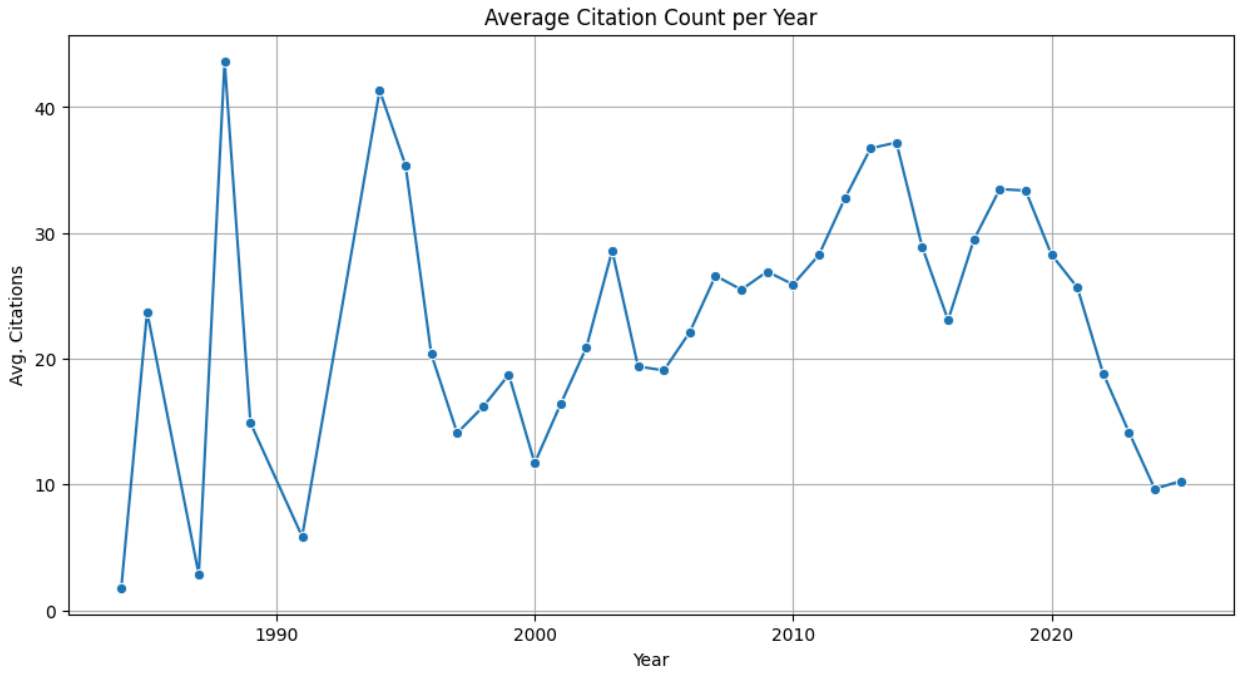
2. **Keyword Popularity Over Time**



3. **Top Author Countries**



4. **Average Citation Count per Year**



---

5. Research Topic Clustering

We used a deep learning-based approach to group similar research papers into thematic clusters.

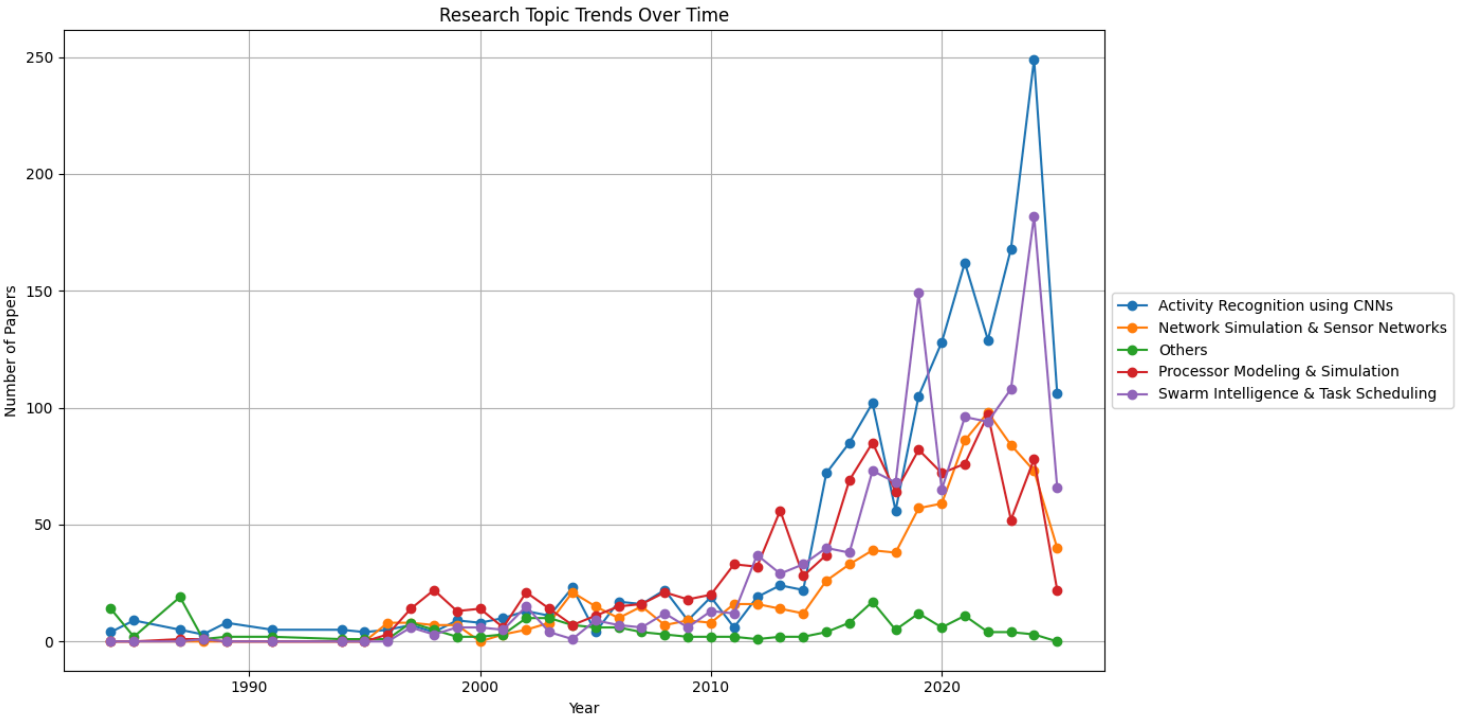
Methodology:

- **Embeddings:** We used **Sentence-BERT** (`all-MiniLM-L6-v2`) to generate semantic embeddings for the abstracts.
- **Clustering:** We applied **KMeans** clustering.
    - The optimal number of clusters (`k=5`) was selected using the **Silhouette Score** method.

- **Topic Labeling:**
    - KeyBERT + manual review were used to label each cluster based on common keywords.

Final Clusters:

```
cluster_labels = {
    0: "Activity Recognition using CNNs",
    1: "Others",  # (Small cluster with mixed topics)
    2: "Swarm Intelligence & Task Scheduling",
    3: "Processor Modeling & Simulation",
    4: "Network Simulation & Sensor Networks"
}
```



---

6. Citation Impact Analysis

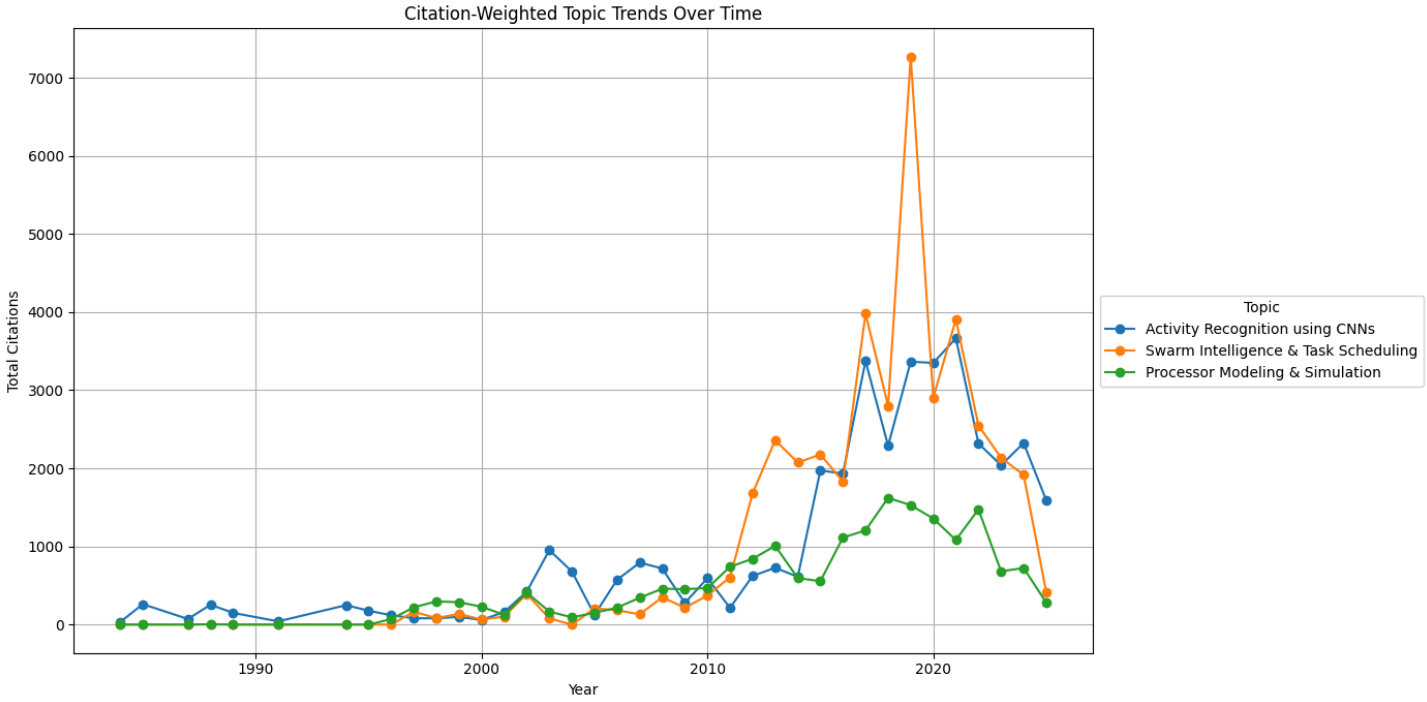
We analyzed how impactful each cluster was based on total and average citations.

- **Citation-Weighted Trends:** Papers were weighted by citation counts rather than only raw counts.

- **Impact Insights:**

    - Activity Recognition topics showed high recent citation growth.

    - Processor Modeling had consistent but moderate citation averages.



---

7. Deep Learning Based Future Trend Prediction

Model Architecture:

- **Model:** LSTM-based Recurrent Neural Network
- **Input:** Each input sample was a sequence of the past 5 years' topic trends (paper counts per topic).
- **Output:** Predicted paper counts per topic for the following year (2026).

Data Preparation:

- The dataset was structured as a time series.

- For each sample, 5 consecutive years' topic counts were used as input, and the paper counts of the immediate next year were set as the target label.
- This method ensures that the model learns temporal patterns across topics.

Technical Details:

- model = Sequential()
- model.add(LSTM(64, activation='relu', input_shape=(5, 5)))
- model.add(Dense(5))

model.compile(optimizer='adam', loss='mse')

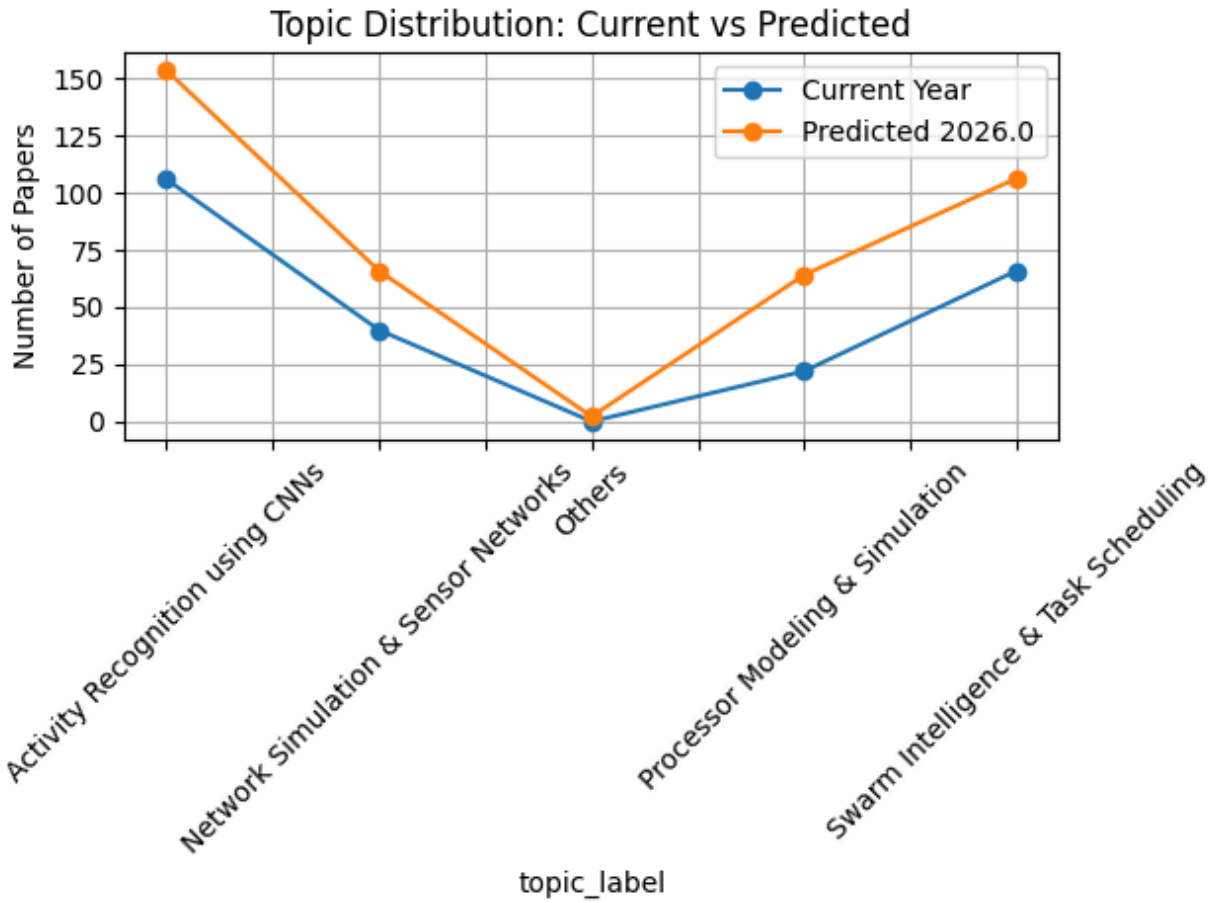
Training Details:

- Loss function: Mean Squared Error (MSE)
- Optimizer: Adam
- Epochs: 100
- TensorFlow version: 2.15.0 (trained on GPU)

The model minimized the MSE loss between the predicted and actual topic distributions year-by-year, learning the underlying sequence patterns across the research topics.

Prediction Year:

- 2026



Observations:

- "Activity Recognition using CNNs" is expected to grow further.
- "Network Simulation & Sensor Networks" maintains steady momentum.
- "Others" category remains unpredictable due to diverse topics.

---

8. Conclusion

- We successfully extracted, cleaned, and analyzed nearly 5000 research papers from 1984 to 2025.

- EDA revealed important patterns in research volume, keyword popularity, and country-wise contributions.

- Using deep learning (Sentence-BERT embeddings + LSTM), we effectively clustered papers into meaningful topics.

- Citation analysis allowed us to evaluate the real-world impact of research areas.

- Our LSTM-based forecasting suggests strong growth in "Activity Recognition using CNNs" and continued relevance for network simulation studies.

Future Work:

- Apply more complex forecasting models (e.g., Transformer-based time series models)

- Improve cluster labeling using advanced NLP techniques (e.g., topic modeling LDA+BERT hybrids)
