



தமிழ்மணம் சர்வதேசத் தமிழ் ஆய்விதழ்

Peer-Reviewed | Open Access | Crossref | Google Scholar
SJIF Impact Factor 3.5 | Multidisciplinary

Article DOI: <https://doi.org/10.63300/tm12022026.45>



Digital Humanities and Tamil Language Resources: AI, OCR, and Digitization of Inscriptions and Palm-Leaf Manuscripts

Dr A.Mohanraj^{1*}, Dr D Mahalingam², Dr.V.C.Srinivasan³,

¹Associate professor in Tamil, Nandha Arts and Science College (Autonomous), Erode -638052

²Head, Department of library & information science

Nandha Arts And Science College.(Autonomous) Erode-52

Email id : nasc.library@gmail.com, **Orchid Id** : <https://orcid.org/0009-0001-3733-9046>

³Administrative Officer, Nandha Arts and Science College (Autonomous), Erode.

Email: srive2345@gmail.com, **Orchid Id**: <https://orcid.org/0009-0008-9926-7982>.

*Corresponding Author Email: drmohanraj1975@gmail.com.

Article Info

Received on 27-June-2026, Revised on 29-June-2026, Accepted on 29-June-2026, Published on 20-July-2026

ABSTRACT

This paper explores the application of Digital Humanities technologies—including Artificial Intelligence (AI), Optical Character Recognition (OCR), and Natural Language Processing (NLP)—to preserve and digitize over two millennia of Tamil linguistic heritage currently threatened by physical decay. It examines the technical challenges of parsing ancient script variations and damaged physical artifacts like palm-leaf manuscripts and stone inscriptions, while evaluating modern computational tools and Tamil-specific Large Language Models (such as Tamil-LLaMA and AI4Tamil). Ultimately, the study highlights how integrating AI-driven workflows safeguards invaluable historical knowledge, advances computational Tamilology, and makes Tamil intellectual heritage globally accessible.

KEYWORDS: Digital Humanities, Tamil Language Resources, AI, OCR, Inscriptions, Palm-leaf Manuscripts, Digitization, Natural Language Processing.



Copyright © 2024 by the author(s). Published by Department of Library, Nallamuthu Gounder Mahalingam College, Pollachi. This is an open access article under the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).



இந்திய அறிவு மரபு (Indian Knowledge Systems - IKS) மற்றும் தமிழ் இலக்கியங்கள்: Viksit Bharat 2047 நோக்கிலான ஆய்வுகள்

¹முனைவர் அ.மோகன்ராஜ், இணைப்பேராசிரியர், தமிழ்த்துறை,

நந்தா கலை மற்றும் அறிவியல் கல்லூரி (தன்னாட்சி), ஈரோடு -638052

²முனைவர் த.மகாலிங்கம், துறைத் தலைவர், நூலகம் மற்றும் தகவல் அறிவியல் துறை,

நந்தா கலை மற்றும் அறிவியல் கல்லூரி (தன்னாட்சி), ஈரோடு 52.

³முனைவர் வி.சி.சீனிவாசன், நிர்வாக அதிகாரி,

நந்தா கலை மற்றும் அறிவியல் கல்லூரி (தன்னாட்சி), ஈரோடு.

சுருக்கம் (Abstract)

டிஜிட்டல் மனிதவியல் (Digital Humanities) என்பது கணினி அறிவியல் தொழில்நுட்பங்களை மனிதநூல் ஆய்வுகளுடன் இணைக்கும் ஒரு புதிய மற்றும் வேகமாக வளர்ந்து வரும் ஆய்வுத்துறையாகும். இது உரைப் பகுப்பாய்வு, தரவுக் காட்சிப்படுத்தல், கணினி உதவி உரை விமர்சனம், மற்றும் டிஜிட்டல் காப்பகங்கள் ஆகியவற்றை உள்ளடக்கியது. தமிழ் மொழி, இரண்டாயிரம் ஆண்டுகளுக்கும் மேலான தொடர்ச்சியான இலக்கிய மரபைக் கொண்டுள்ளது. இந்தப் பரந்த இலக்கியத் தொகுப்பு, மொழியியல், சமூகவியல், வரலாறு, தத்துவம் ஆகிய பல்வேறு தளங்களில் அரிய அறிவினை வழங்குகின்றது.

எனினும், ஓலைச்சுவடிகள், கல்வெட்டுகள், மற்றும் பழைய அச்சுப்பதிப்புகள் வடிவில் பரந்த அளவிலான தமிழ் மொழி வளங்கள் தற்போது பாதுகாக்கப்படாமல் உள்ளன. ஈரப்பதம், பூச்சிகள், மற்றும் காலப் போக்கில் ஏற்படும் அரிப்பு ஆகியவற்றால் இந்த அரிய பதிவுகள் படிப்படியாக அழிந்து வருகின்றன. பல நூற்றாண்டுகளின் அறிவைத் தன்னகத்தே கொண்ட இவ்வளங்களைப் பாதுகாப்பதும், அணுகும்படி செய்வதும் காலத்தின் கட்டாயத் தேவையாகும்.

இக்கட்டுரை, டிஜிட்டல் மனிதவியல் கருவிகளைப் பயன்படுத்தி தமிழ் மொழி வளங்களை மின்னாக்கம் செய்வதற்கான வாய்ப்புகளையும், சவால்களையும் ஆழமாக ஆராய்கிறது. செயற்கை நுண்ணறிவு (AI) தொழில்நுட்பங்கள், ஒளிவாங்கி எழுத்துணர்வு (OCR), மற்றும் இயற்கை மொழி செயலாக்கம் (NLP) ஆகியவை தமிழ் கல்வெட்டுகள் மற்றும் ஓலைச்சுவடிகளை டிஜிட்டல் மயமாக்குவதில் எவ்வாறு பயன்படுத்தப்படலாம் என்பதை விரிவாக விவரிக்கிறது. குறிப்பாக, எழுத்து வடிவங்களின் பன்முகத்தன்மை, மங்கிய எழுத்துக்களை அடையாளம் காணும் OCR சவால்கள், மற்றும் கையெழுத்துப் பிரதிகளை பகுப்பாய்வு செய்வதில் AI-இன் பங்கு ஆகியவை விவாதிக்கப்படுகின்றன.

மேலும், தமிழுக்கான கணினித் தமிழியல் உள்கட்டமைப்பின் தற்போதைய நிலையும் இக்கட்டுரையில் மதிப்பிடப்படுகிறது. Tamil-LLaMA, AI4Tamil, மற்றும் DravidianCodeMix போன்ற தமிழ்-குறிப்பிட்ட பெருமொழி மாதிரிகள் (LLMs) மற்றும் தரவுத்தொகுப்புகள், தமிழ் மொழியியல் ஆராய்ச்சியை எவ்வாறு மேம்படுத்துகின்றன என்பது ஆராயப்படுகிறது. இந்தக் கருவிகள் பெருமளவிலான உரைத் தொகுப்புகளை விரைவாக பகுப்பாய்வு செய்வதற்கும், புதிய தகவல்களை வெளிக்கொணர்வதற்கும் உதவுகின்றன.

இறுதியாக, இத்துறையில் உள்ள தொழில்நுட்ப வரம்புகள், நெறிமுறைக் கவலைகள், மற்றும் எதிர்கால ஆராய்ச்சி திசைகள் பற்றியும் இக்கட்டுரை விவாதிக்கிறது. டிஜிட்டல் மனிதவியல் கருவிகள், தமிழ் மொழி வளங்களைப் பாதுகாத்து, எதிர்கால சந்ததியினருக்கு அணுகும் வாய்ப்பை வழங்குவதோடு, உலகளவில் தமிழ் அறிவு மரபைப் பகிர்ந்துகொள்ளவும் உதவும் என்பதில் ஐயமில்லை.

முக்கியச் சொற்கள்: டிஜிட்டல் மனிதவியல், தமிழ் மொழி வளங்கள், AI, OCR, கல்வெட்டுகள், ஓலைச்சுவடிகள், மின்னாக்கம், இயற்கை மொழி செயலாக்கம்

1. முன்னுரை

டிஜிட்டல் மனிதவியல் (Digital Humanities - DH) என்பது கணினி அறிவியல் தொழில்நுட்பங்களை மனிதநூல் ஆய்வுகளுடன் இணைக்கும் ஒரு புதிய மற்றும் வேகமாக வளர்ந்து வரும் ஆய்வுத்துறையாகும். இது உரைப் பகுப்பாய்வு, தரவுக் காட்சிப்படுத்தல், கணினி உதவி உரை விமர்சனம், மற்றும் டிஜிட்டல் காப்பகங்கள் ஆகியவற்றை உள்ளடக்கியது (Svensson, 2016; Schreibman et al., 2020). டிஜிட்டல் மனிதவியல், பாரம்பரிய ஆய்வு முறைகளுக்கு மாற்றாக அல்ல, மாறாக அவற்றை நிறைவு செய்யும் கருவிகளாக செயல்படுகிறது.

தமிழ் மொழி, இரண்டாயிரம் ஆண்டுகளுக்கும் மேலான தொடர்ச்சியான இலக்கிய மரபைக் கொண்டுள்ளது. சங்க இலக்கியம் முதல் நவீன இலக்கியம் வரையிலான இந்தப் பரந்த தொகுப்பு, மொழியியல், சமூகவியல், வரலாறு, தத்துவம் ஆகிய பல்வேறு தளங்களில் ஆய்வுக்கு உட்படுத்தப்பட்டு வருகிறது. எனினும், இந்த வளங்களில் பெரும்பகுதி ஓலைச்சுவடிகள், கல்வெட்டுகள், மற்றும் பழைய அச்சுப்பதிப்புகள் வடிவில் பாதுகாக்கப்படாமல் உள்ளது (Balachandran, 2023; Mohan & Vijayakumar, 2022). இந்த அரிய பதிவுகள், ஈரப்பதம், பூச்சிகள், காலப்போக்கில் ஏற்படும் இயற்கைச் சிதைவு போன்றவற்றால் படிப்படியாக அழிந்து வருகின்றன. இந்த நிலையில், டிஜிட்டல் மனிதவியல் கருவிகள் தமிழ் மொழி வளங்களை பாதுகாத்து, ஆய்வு செய்து, பகிர்ந்துகொள்ளும் புதிய வாய்ப்புகளைத் திறக்கின்றன. செயற்கை நுண்ணறிவு (AI),

ஒளிவாங்கி எழுத்துணர்வு (OCR), இயற்கை மொழி செயலாக்கம் (NLP) ஆகிய தொழில்நுட்பங்கள் தமிழ் கல்வெட்டுகள் மற்றும் ஓலைச்சுவடிகளை டிஜிட்டல் மயமாக்குவதில் முக்கிய பங்கு வகிக்கின்றன (Chakravarthi et al., 2022; Doddapaneni et al., 2023). இந்தக் கருவிகள் பெருமளவிலான உரைத் தொகுப்புகளை விரைவாக பகுப்பாய்வு செய்வதற்கும், புதிய தகவல்களை வெளிக்கொணர்வதற்கும் உதவுகின்றன.

இக்கட்டுரை, டிஜிட்டல் மனிதவியல் கருவிகளைப் பயன்படுத்தி தமிழ் மொழி வளங்களை மின்னாக்கம் செய்வதற்கான வாய்ப்புகளையும், சவால்களையும் ஆராய்கிறது. குறிப்பாக, AI, OCR, கல்வெட்டுகள், மற்றும் ஓலைச்சுவடிகளின் மின்னாக்கம் ஆகியவற்றின் தற்போதைய நிலை, எதிர்கால சாத்தியக்கூறுகள், மற்றும் நெறிமுறைக் கவலைகள் பற்றி விவாதிக்கிறது.

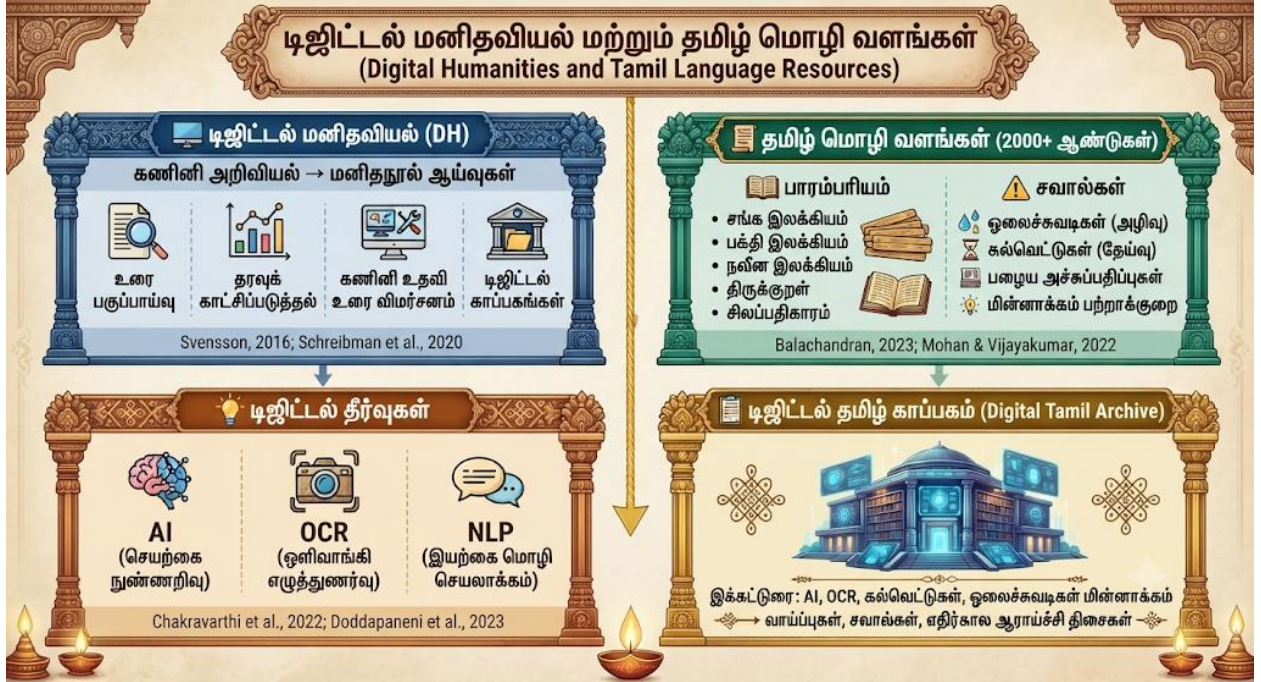


Fig 1: டிஜிட்டல் மனிதவியல் மற்றும் தமிழ் மொழி வளங்கள்

இந்தப் படம் டிஜிட்டல் மனிதவியல் தொழில்நுட்பங்களை தமிழ் மொழி வளங்களுடன் இணைப்பதை விளக்குகிறது. கணினி அறிவியல் மற்றும் செயற்கை நுண்ணறிவு (AI), OCR, NLP போன்ற கருவிகள் மூலம் ஓலைச்சுவடிகள் மற்றும் பாரம்பரிய தமிழ் இலக்கியங்களை மின்னாக்கம் செய்து பாதுகாப்பதற்கான வழிமுறைகளையும் காப்பகங்களையும் இது தெளிவுபடுத்துகிறது.

2. டிஜிட்டல் மனிதவியல்: ஒரு கண்ணோட்டம்

டிஜிட்டல் மனிதவியல் என்பது பல்துறை ஆய்வுக் களமாகும். இது கணினி அறிவியல் கருவிகளைப் பயன்படுத்தி மனிதநூல் ஆய்வுகளை மேற்கொள்கிறது (Schreibman et al., 2020). இது உரைப்

பகுப்பாய்வு, தரவுக் காட்சிப்படுத்தல், இடஞ்சார் பகுப்பாய்வு, நெட்வொர்க் பகுப்பாய்வு, மற்றும் டிஜிட்டல் காப்பகங்கள் ஆகியவற்றை உள்ளடக்கியது.

டிஜிட்டல் மனிதவியலின் முக்கிய நோக்கங்கள்:

- **காப்புப் பாதுகாப்பு:** பழைய, உடையக்கூடிய ஆவணங்களை டிஜிட்டல் மயமாக்கி பாதுகாப்பது. இது எதிர்கால சந்ததியினருக்கு இந்த அரிய பதிவுகளை அணுகும் வாய்ப்பை வழங்குகிறது.
- **அணுகுமை:** பரந்த பொதுமக்களுக்கும், ஆய்வாளர்களுக்கும் வளங்களை அணுகச் செய்வது. இணையவழி காப்பகங்கள் மூலம் உலகெங்கிலும் உள்ள ஆய்வாளர்கள் தமிழ் வளங்களைப் பயன்படுத்தலாம்.
- **பகுப்பாய்வு:** பெருமளவிலான உரைத் தொகுப்புகளை விரைவாக பகுப்பாய்வு செய்து புதிய கண்டுபிடிப்புகளை உருவாக்குவது. AI கருவிகள் மனித ஆய்வாளர்களால் சாத்தியமில்லாத விரைவில் பெரிய தரவுகளை பகுப்பாய்வு செய்கின்றன.
- **காட்சிப்படுத்தல்:** சிக்கலான தரவுகளை காட்சி வடிவில் முன்வைத்து புரிந்துகொள்ள உதவுவது. இது புதிய உள்ளுணர்வுகளை உருவாக்க உதவுகிறது.
- **ஒத்துழைப்பு:** பல்துறை ஆய்வுக் குழுக்கள் இணைந்து பணியாற்றுவதை ஊக்குவிப்பது. இலக்கிய ஆய்வாளர்கள், வரலாற்றாசிரியர்கள், கணினி விஞ்ஞானிகள் இணைந்து பணியாற்ற உதவுகிறது (Svensson, 2016).

2.2 டிஜிட்டல் மனிதவியலின் வளர்ச்சி

டிஜிட்டல் மனிதவியல் துறை 1940-களில் ராபர்டோ பூசாவின் இலக்கியக் கணக்கீட்டு ஆய்வுகளுடன் தொடங்கியது. 1960-70-களில் உரைப் பகுப்பாய்வு மற்றும் கணினி உதவி உரை விமர்சனம் வளர்ந்தன. 1990-களில் இணையத்தின் வளர்ச்சியுடன் டிஜிட்டல் காப்பகங்கள் உருவாகின. 2000-களில் இருந்து, AI மற்றும் பெருந்தரவு (Big Data) தொழில்நுட்பங்கள் டிஜிட்டல் மனிதவியலில் புதிய பரிமாணங்களைத் திறந்துள்ளன (Schreibman et al., 2020).

இந்தியாவில், டிஜிட்டல் மனிதவியல் துறை ஒப்பீட்டளவில் புதியது. எனினும், சமீப காலங்களில் இந்திய மொழிகளுக்கான கணினித் தமிழியல் தரவுத்தொகுப்புகள் மற்றும் மொழி மாதிரிகள் உருவாக்கப்பட்டு வருகின்றன (Doddapaneni et al., 2023). இவை தமிழ் இலக்கிய ஆய்வுக்கு புதிய வாய்ப்புகளைத் திறக்கின்றன.

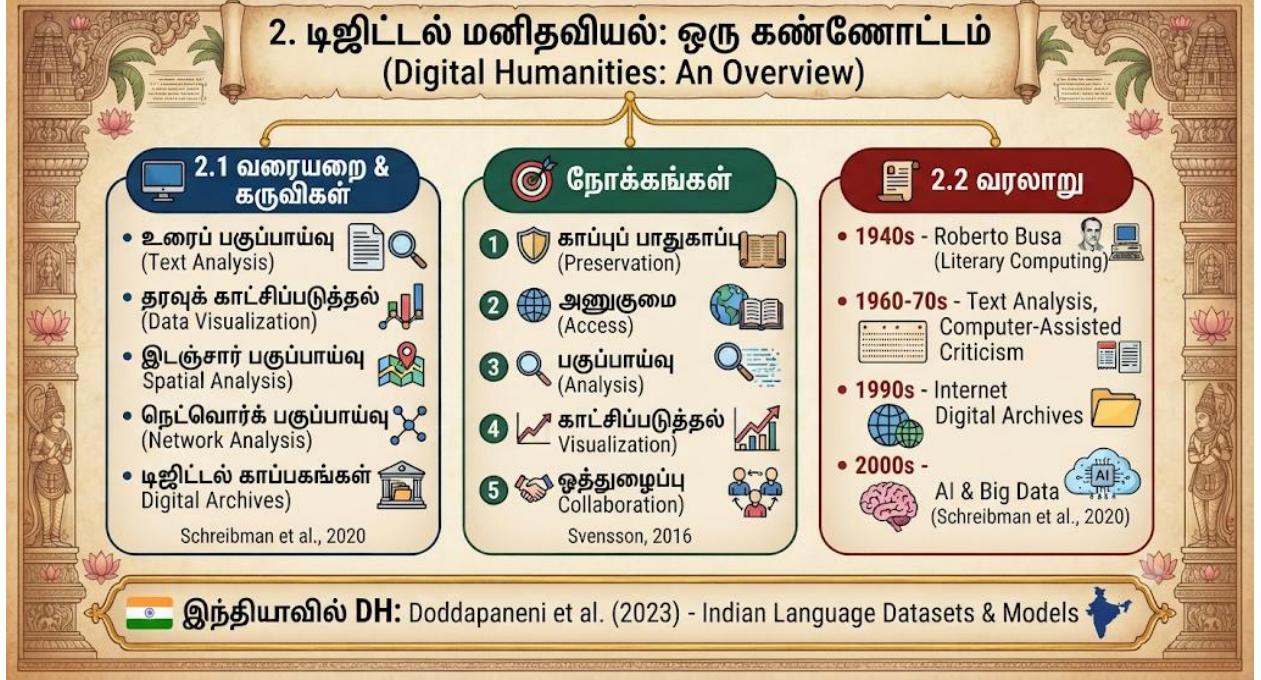


Fig 2: டிஜிட்டல் மனிதவியல்: ஒரு கண்ணோட்டம்

இந்தப் படம் டிஜிட்டல் மனிதவியல் (Digital Humanities) பற்றிய ஒரு கண்ணோட்டத்தை வழங்குகிறது. இது மூன்று முக்கிய பகுதிகளைக் கொண்டுள்ளது:

1. **வரையறை & கருவிகள்:** உரைப் பகுப்பாய்வு, தரவுக் காட்சிப்படுத்தல் மற்றும் நெட்வொர்க் பகுப்பாய்வு போன்ற டிஜிட்டல் கருவிகளை விளக்குகிறது.
2. **நோக்கங்கள்:** கலாச்சாரப் பொருட்களைப் பாதுகாத்தல், அணுகலை எளிதாக்குதல் மற்றும் கூட்டு ஆராய்ச்சியை ஊக்குவித்தல் ஆகிய இதன் முக்கிய நோக்கங்களை பட்டியலிடுகிறது.
3. **வரலாறு:** 1940களில் ரொபர்டோ பூசாவின் இலக்கியக் கம்ப்யூட்டிங் முதல் நவீன AI மற்றும் பெரிய தரவு (Big Data) பயன்பாடு வரையிலான அதன் பரிணாமத்தை விவரிக்கிறது.

கூடுதலாக, இது இந்தியாவில் டிஜிட்டல் மனிதவியலின் தற்போதைய நிலையையும், இந்திய மொழிக் தரவுத்தொகுப்புகள் மற்றும் மாதிரிகளின் முக்கியத்துவத்தையும் எடுத்துக்காட்டுகிறது.

3. தமிழ் மொழி வளங்கள்: தற்போதைய நிலை

3.1 ஓலைச்சுவடிகள்

தமிழ் மொழியின் பழமையான ஓலைச்சுவடிகள் கி.பி. 1-2 நூற்றாண்டுகளைச் சேர்ந்தவை. இவை தமிழ் இலக்கியம், இலக்கணம், மருத்துவம், ஜோதிடம், கணிதம், மற்றும் தத்துவம் ஆகிய பல்வேறு துறைகளின் அறிவைப் பதிவு செய்துள்ளன (Balachandran, 2023). இந்த ஓலைச்சுவடிகள் மிகவும் உடையக்கூடியவை; பூச்சிகள், ஈரப்பதம், மற்றும் பராமரிப்பின்மை காரணமாக பல சுவடிகள் அழிந்து வருகின்றன.

தமிழகத்தில் தற்போது சுமார் 50,000 ஓலைச்சுவடிகள் பாதுகாக்கப்பட்டு வருகின்றன. இவற்றில் பெரும்பகுதி தமிழ்ப் பல்கலைக்கழகம், சென்னை மாநில நூலகம், மற்றும் பல அருங்காட்சியகங்களில் உள்ளது. எனினும், இந்த சுவடிகளில் பெரும்பகுதி இன்னும் மின்னாக்கம் செய்யப்படவில்லை (Mohan & Vijayakumar, 2022).

3.2 கல்வெட்டுகள்

தமிழ் கல்வெட்டுகள் கி.மு. 3-ம் நூற்றாண்டு முதல் தொடர்கின்றன. இவை தமிழ் பிராமி, வட்டெழுத்து, கிரந்தம், மற்றும் தமிழ் எழுத்துக்களில் பொறிக்கப்பட்டுள்ளன. தமிழகத்தில் தற்போது 25,000 கல்வெட்டுகள் பதிவு செய்யப்பட்டுள்ளன (Rangarajan, 2017). இவை அரசியல், வரலாறு, பொருளியல், மதம், மற்றும் சமூக வாழ்க்கை பற்றிய முக்கிய தகவல்களை வழங்குகின்றன.

3.3 பழைய அச்சுப்பதிப்புகள்

தமிழில் முதல் அச்சுப் புத்தகம் 1578-இல் பதிப்பிக்கப்பட்டது. இதிலிருந்து பல ஆயிரம் தமிழ் நூல்கள் வெளிவந்துள்ளன. இவற்றில் பல இன்று அரிய நூல்களாக உள்ளன; அவை அச்சிலிருந்து மின்னாக்கம் செய்யப்பட வேண்டும்.

3.4 கணினித் தமிழியல் உள்கட்டமைப்பு

தமிழுக்கான கணினித் தமிழியல் உள்கட்டமைப்பு படிப்படியாக வலுப்படுத்தப்பட்டு வருகிறது. Tamil-LLaMA, AI4Tamil, மற்றும் DravidianCodeMix போன்ற தமிழ்-குறிப்பிட்ட மொழி மாதிரிகள் உருவாக்கப்பட்டுள்ளன (Chakravarthi et al., 2022; Doddapaneni et al., 2023). இவை தமிழ் இலக்கிய ஆய்வுக்கு புதிய வாய்ப்புகளைத் திறக்கின்றன.



Fig 3: தமிழ் மொழி வளங்கள்: தற்போதைய நிலை மற்றும் எதிர்காலம்

இப்படம் தமிழ் மொழி வளங்களின் தற்போதைய நிலையை விளக்குகிறது. இது ஓலைச்சுவடிகள், கல்வெட்டுகள் மற்றும் பழைய அச்சுப் பதிப்புகள் ஆகியவற்றின் முக்கியத்துவத்தையும், அவை எதிர்கொள்ளும் சவால்களையும் (பூச்சிகள், தேய்மானம்) பட்டியலிடுகிறது. மேலும், தமிழ் மொழிக்கான நவீன கணினி உள்கட்டமைப்பு (Tamil-LLaMA போன்றவை) மற்றும் புதிய ஆராய்ச்சி வாய்ப்புகள் பற்றியும் இது எடுத்துரைக்கிறது.

4. AI மற்றும் OCR: தமிழ் மொழி வளங்களின் மின்னாக்கம்

4.1 OCR தொழில்நுட்பம்

ஒளிவாங்கி எழுத்துணர்வு (Optical Character Recognition - OCR) என்பது படிமங்களில் உள்ள எழுத்துகளை இயந்திரத்தால் படிக்கக்கூடிய உரையாக மாற்றும் தொழில்நுட்பமாகும். தமிழ் OCR, பழைய அச்சுப்பதிப்புகள், ஓலைச்சுவடிகள், மற்றும் கையெழுத்துப் பிரதிகளை மின்னாக்கம் செய்வதில் முக்கிய பங்கு வகிக்கிறது. இந்த தொழில்நுட்பம் பாரம்பரிய ஆவணங்களை டிஜிட்டல் வடிவத்திற்கு மாற்றி, அவற்றை பாதுகாத்து, எளிதாக ஆய்வு செய்ய உதவுகிறது.

தமிழ் OCR-இன் முக்கிய சவால்கள்:

1. **எழுத்து வடிவங்களின் பன்முகத்தன்மை:** தமிழில் அச்சு, கையெழுத்து, கல்வெட்டு எழுத்துகள் என பல வடிவங்கள் உள்ளன. இவை ஒவ்வொன்றும் வேறுபட்ட அமைப்பைக் கொண்டிருப்பதால், OCR இயந்திரங்கள் அவற்றை அடையாளம் காண்பதில் சிரமப்படுகின்றன.

2. எழுத்து இணைப்புகளின் சிக்கல்: தமிழில் உயிர்மெய் எழுத்துகள் இணைந்து வருவதால் (எ.கா., க+அ = க, க+ஆ = கா), ஒற்றை எழுத்துகளை இயந்திரம் பிரித்தறிவது கடினமாகிறது. இந்த இணைப்புகள் OCR துல்லியத்தை கணிசமாக பாதிக்கின்றன.
3. ஓலைச்சுவடி சிராய்ப்பு: பழைய ஓலைச்சுவடிகளில் எழுத்துகள் மங்கி அல்லது சிராய்ந்து இருப்பதால், அவற்றை OCR மூலம் படிப்பது மிகவும் சவாலானதாகும். இயற்கைச் சிதைவு, ஈரப்பதம், பூச்சிகள் போன்றவை எழுத்துகளை அழித்துவிடுகின்றன.
4. கையெழுத்து வேறுபாடுகள்: ஒவ்வொரு எழுத்தாளரின் கையெழுத்தும் வேறுபடுவதால், OCR மாதிரிகள் ஒரு சீரான வடிவத்தை அடையாளம் காண இயலாமல் தடுமாறுகின்றன. இது கையெழுத்துப் பிரதிகளின் மின்னாக்கத்தை கடினமாக்குகிறது (Mohan & Vijayakumar, 2022).

4.2 AI-இன் பங்கு

செயற்கை நுண்ணறிவு (AI) மற்றும் இயந்திர கற்றல் (Machine Learning) தொழில்நுட்பங்கள் OCR-இன் துல்லியத்தை கணிசமாக மேம்படுத்தியுள்ளன. குறிப்பாக, ஆழ்கற்றல் (Deep Learning) நுட்பங்கள் தமிழ் OCR-இன் துல்லியத்தை 90% வரை உயர்த்தியுள்ளன (Ganesan & Krishnamurthy, 2024).

1. AI, OCR-இல் பின்வரும் வழிகளில் உதவுகிறது:
2. எழுத்து அடையாளங்காணல்: AI மாதிரிகள் பல்வேறு எழுத்து வடிவங்களையும், கையெழுத்து மாறுபாடுகளையும் அடையாளம் காண பயிற்சி பெறுகின்றன. இவை பெருமளவிலான தரவுகளைக் கொண்டு பயிற்சியளிக்கப்பட்டு, துல்லியமான முடிவுகளை வழங்குகின்றன.
3. பிழை திருத்தம்: AI, OCR-இல் ஏற்படும் பிழைகளை தானாக கண்டுபிடித்து திருத்துகிறது. இது மேற்கொள்ளப்படும் உரை துல்லியமாக இருப்பதை உறுதி செய்கிறது.
4. உரை மறுசீரமைப்பு: AI, சிதைந்த அல்லது மங்கிய எழுத்துகளை மறுசீரமைக்க உதவுகிறது. எழுத்துகளின் அமைப்பு, வடிவம், மற்றும் சூழல் ஆகியவற்றை ஆராய்ந்து, இழந்த பகுதிகளை ஊகித்து நிரப்புகிறது.
5. கையெழுத்து அடையாளங்காணல்: AI, கையெழுத்துப் பிரதிகளை படிக்கும் திறனை மேம்படுத்துகிறது. இது பல்வேறு கையெழுத்து வடிவங்களை அடையாளம் காணும் திறனைப் பெற்று, துல்லியமாக உரையாக மாற்றுகிறது (Sahoo et al., 2024).

இவ்வாறு, AI மற்றும் OCR இணைந்து தமிழ் மொழி வளங்களின் மின்னாக்கத்தில் புதிய பரிமாணங்களை உருவாக்குகின்றன. இத்தொழில்நுட்பங்கள் பழம்பெரும் தமிழ் ஆவணங்களை எதிர்கால சந்ததியினருக்கு பாதுகாக்க உதவுகின்றன.



Fig4: தமிழ் மொழி வள மின்னாக்கத்தில் OCR மற்றும் செயற்கை நுண்ணறிவு (AI)

இப்பதிவு தமிழ் மொழி வளங்களை மின்னாக்கம் செய்வதில் OCR மற்றும் செயற்கை நுண்ணறிவு (AI) ஆகியவற்றின் முக்கியத்துவத்தை விளக்குகிறது. தமிழ் எழுத்து வடிவங்களின் பன்முகத்தன்மை, எழுத்து இணைப்புகள் மற்றும் ஒலைச்சுவடி சிராய்ப்புகள் போன்ற OCR-இன் சவால்களை இது பட்டியலிடுகிறது. மேலும், ஆழ்கற்றல் நுட்பங்கள் மூலம் துல்லியத்தை 90% வரை உயர்த்துவதுடன், எழுத்து அடையாளங்காணல் மற்றும் பிழை திருத்தங்களில் AI எவ்வாறு உதவுகிறது என்பதையும் விவரிக்கிறது.

5. கல்வெட்டுகளின் மின்னாக்கம்

5.1 கல்வெட்டு OCR

கல்வெட்டு OCR என்பது கல்வெட்டுகளில் உள்ள எழுத்துகளை இயந்திரத்தால் படிக்கக்கூடிய உரையாக மாற்றும் தொழில்நுட்பமாகும். பழங்காலக் கற்பாறைகளில் பொறிக்கப்பட்டுள்ள இந்த வரலாற்றுப் பதிவுகள், அரசியல், பொருளியல், மதம், மற்றும் சமூக வாழ்க்கை பற்றிய முக்கிய தகவல்களை வழங்குகின்றன. இவற்றை மின்னாக்கம் செய்வதன் மூலம், அவற்றை பாதுகாத்து, எளிதாக ஆய்வு செய்யவும், பகிர்ந்துகொள்ளவும் முடியும்.

தமிழ் கல்வெட்டுகளின் OCR-இன் தனித்துவமான சவால்கள்:

1. **கற்பாறை மேற்பரப்பின் கடினத்தன்மை:** கற்பாறைகளின் கடினமான, கரடுமுரடான மேற்பரப்பில் எழுத்துகள் பொறிக்கப்பட்டுள்ளன. இவற்றை உயர்தர படிமங்களாகப் படமாக்குவதும், எழுத்துகளை தெளிவாக அடையாளம் காண்பதும் மிகவும் கடினமானதாகும். ஒளி மற்றும் நிழலின் தாக்கம் காரணமாக எழுத்துகள் சிதைந்து தெரியும்.
2. **எழுத்து அரிப்பு:** பல நூற்றாண்டுகள் காற்று, மழை, மற்றும் வெப்பநிலை மாற்றங்களுக்கு உட்பட்ட கல்வெட்டுகளில் எழுத்துகள் படிப்படியாக அரிக்கப்பட்டு மங்கிப்போயின. இதனால், பல எழுத்துகளை OCR இயந்திரங்களால் அடையாளம் காண இயலாமல் போகிறது.
3. **வெவ்வேறு எழுத்து முறைகளின் பன்முகத்தன்மை:** தமிழ் கல்வெட்டுகள் தமிழ் பிராமி (கி.மு. 3-ம் நூற்றாண்டு), வட்டெழுத்து (கி.பி. 5-12ம் நூற்றாண்டு), கிரந்தம் (சமஸ்கிருதத்திற்கு), மற்றும் தற்கால தமிழ் எழுத்துகள் என பல்வேறு முறைகளில் பொறிக்கப்பட்டுள்ளன. ஒவ்வொரு முறைக்கும் தனித்துவமான எழுத்து வடிவங்கள் உள்ளன.
4. **மொழி வேறுபாடுகள்:** கல்வெட்டுகள் தமிழ் மட்டுமின்றி, சமஸ்கிருதம், கன்னடம், தெலுங்கு போன்ற பல மொழிகளில் இடம்பெற்றுள்ளன. இந்த பன்மொழி உள்ளடக்கத்தை கையாளும் OCR அமைப்புகள் மிகவும் சிக்கலானவை (Rangarajan, 2017).

5.2 AI-உதவி கல்வெட்டு பகுப்பாய்வு

செயற்கை நுண்ணறிவு (AI) கருவிகள் கல்வெட்டுகளின் பகுப்பாய்வில் புதிய பரிமாணங்களைத் திறக்கின்றன (Veerakannan & Vijayakumar, 2025):

1. **எழுத்து அடையாளங்காணல்:** AI மாதிரிகள், ஆயிரக்கணக்கான கல்வெட்டுப் படிமங்களில் பயிற்சி பெற்று, பல்வேறு வடிவங்களிலான பழங்கால எழுத்துகளை துல்லியமாக அடையாளம் காண்கின்றன.
2. **காலம் கணிப்பு:** AI, எழுத்துகளின் வடிவம், பாணி, மற்றும் அமைப்பு ஆகியவற்றை வைத்து, கல்வெட்டு பொறிக்கப்பட்ட தோராயமான காலத்தை கணிக்க உதவுகிறது. இது வரலாற்று நிகழ்வுகளை காலவரிசைப்படுத்த உதவுகிறது.
3. **உரை மறுசீரமைப்பு:** AI, சிதைந்த, மங்கிய, அல்லது உடைந்த எழுத்துகளின் சூழலை ஆராய்ந்து, காணாமல் போன பகுதிகளை ஊகித்து உரையை மறுசீரமைக்க உதவுகிறது.
4. **டிஜிட்டல் தரவுத்தள உருவாக்கம்:** AI, கல்வெட்டுகளை டிஜிட்டல் காப்பகங்களாக மாற்றி, அவற்றை தேடக்கூடிய, பகுப்பாய்வு செய்யக்கூடிய தரவுத்தளங்களாக உருவாக்க

உதவுகிறது (Chakravarthi et al., 2022). இது உலகெங்கிலும் உள்ள ஆய்வாளர்களுக்கு இந்த அரிய வளங்களை அணுகும் வாய்ப்பை வழங்குகிறது.



Fig5: கல்வெட்டுகளின் மின்னாக்கம்: சவால்களும் AI தீர்வுகளும்

இந்தப் படம் தமிழ் கல்வெட்டுகளை டிஜிட்டல் மயமாக்கும் செயல்முறையை விளக்குகிறது. இது கல்வெட்டுகளை மின்னூரையாக மாற்றும் OCR தொழில்நுட்பம் (Rangarajan, 2017), கற்பாறை மேற்பரப்பு மற்றும் அரிப்பு போன்ற சவால்கள், மற்றும் AI-உதவி பகுப்பாய்வு (Veerakannan & Vijayakumar, 2025) ஆகியவற்றை உள்ளடக்கியது. ஆழ்கற்றல் (Deep Learning) தொழில்நுட்பம் பன்மொழி உள்ளடக்கத்தை துல்லியமாக பகுப்பாய்வு செய்வதில் புதிய வாய்ப்புகளை ஏற்படுத்துகிறது.

6. ஓலைச்சுவடிகளின் மின்னாக்கம்

6.1 ஓலைச்சுவடி OCR

ஓலைச்சுவடி OCR, பழைய ஓலைச்சுவடிகளை மின்னாக்கம் செய்யும் தொழில்நுட்பமாகும். இது கல்வெட்டு OCR-ஐ விட கூடுதல் சவால்களைக் கொண்டுள்ளது. ஓலைச்சுவடிகள் தமிழ் மொழியின் பழமையான பதிவுகளைக் கொண்டுள்ளன; இவற்றைப் பாதுகாப்பது மிகவும் முக்கியமானதாகும்.

ஓலைச்சுவடி OCR-இன் முக்கிய சவால்கள்:

1. ஓலை மேற்பரப்பின் தனித்தன்மை: ஓலைகளின் மென்மையான, ஆனால் சீரற்ற மேற்பரப்பில் எழுத்துகள் கூரான எழுத்தாணி மூலம் பொறிக்கப்பட்டுள்ளன. இவற்றை உயர்தர படிமங்களாகப் படமாக்குவது கடினம்; ஒளி மற்றும் நிழலின் தாக்கம் எழுத்துகளை சிதைக்கும்.
2. எழுத்து மங்கல்: பல நூற்றாண்டுகள் பழமையான ஓலைச்சுவடிகளில் எழுத்துகள் படிப்படியாக மங்கி உள்ளன. இயற்கையான காலப்போக்கு, ஈரப்பதம், மற்றும் பயன்பாடு காரணமாக எழுத்துகளின் தெளிவு குறைந்துவிடுகிறது.
3. ஓலை சிதைவு: ஓலைச்சுவடிகள் பூச்சிகள் (குறிப்பாக கரையான்கள்), ஈரப்பதம், மற்றும் பராமரிப்பின்மை காரணமாக பலவீனமடைந்து, உடைந்து, அல்லது முற்றிலும் அழிந்து விடுகின்றன. இது OCR செயல்முறையை மிகவும் சிக்கலாக்குகிறது.
4. கையெழுத்து வேறுபாடுகள்: ஒவ்வொரு எழுத்தாளரின் கையெழுத்தும் தனித்துவமானது; எழுத்து அளவு, சாய்வு, அழுத்தம், மற்றும் வடிவத்தில் வேறுபாடுகள் உள்ளன. OCR இயந்திரங்கள் இந்த பன்முகத்தன்மையைக் கையாள்வதில் சிரமப்படுகின்றன (Balachandran, 2023).

6.2 ஓலைச்சுவடி காப்பகங்கள்

தமிழ் ஓலைச்சுவடிகளின் முக்கிய காப்பகங்கள்:

1. தமிழ்ப் பல்கலைக்கழகம், தஞ்சாவூர்: 20,000 க்கும் மேற்பட்ட ஓலைச்சுவடிகள், இவற்றில் பல அரிய இலக்கியப் படைப்புகள் உள்ளன.
2. சென்னை மாநில நூலகம்: 10,000 க்கும் மேற்பட்ட ஓலைச்சுவடிகள், கடலோரத் தமிழகத்தின் பதிவுகளைக் கொண்டவை.
3. கேரள அருங்காட்சியகம், திருவனந்தபுரம்: தமிழ் மற்றும் மலையாள ஓலைச்சுவடிகள், இரண்டு மொழிகளுக்கிடையிலான தொடர்புகளை வெளிப்படுத்துபவை.
4. பிரிட்டிஷ் நூலகம், இலண்டன்: காலனித்துவ காலத்தில் சேகரிக்கப்பட்ட தமிழ் ஓலைச்சுவடிகள்.
5. பாரிஸ் கிழக்கத்திய நூலகம்: பிரெஞ்சு அறிஞர்களால் சேகரிக்கப்பட்ட தமிழ் ஓலைச்சுவடிகள் (Mohan & Vijayakumar, 2022).

6.3 AI-உதவி ஓலைச்சுவடி பகுப்பாய்வு

1. AI கருவிகள் ஓலைச்சுவடிகளின் பகுப்பாய்வில் புதிய வாய்ப்புகளைத் திறக்கின்றன (Veerakannan & Vijayakumar, 2025):
2. எழுத்து அடையாளங்காணல்: AI, பல்வேறு கையெழுத்து வடிவங்களை அடையாளம் காண பயிற்சி பெற்று, துல்லியமாக எழுத்துகளை அடையாளம் காட்டுகிறது.
3. உரை மறுசீரமைப்பு: AI, சிதைந்த, உடைந்த, அல்லது மங்கிய பகுதிகளின் குழலை ஆராய்ந்து, காணாமல் போன உரைப் பகுதிகளை ஊகித்து மறுசீரமைக்க உதவுகிறது.
4. உள்ளடக்க பகுப்பாய்வு: AI, ஓலைச்சுவடிகளின் உள்ளடக்கத்தைப் பகுப்பாய்வு செய்து, கருப்பொருள்கள், தலைப்புகள், மற்றும் முக்கிய கருத்துகளை அடையாளம் காண்கிறது.
5. டிஜிட்டல் தரவுத்தள உருவாக்கம்: AI, ஓலைச்சுவடிகளின் தேடக்கூடிய, பகுப்பாய்வு செய்யக்கூடிய டிஜிட்டல் காப்பகங்களை உருவாக்க உதவுகிறது (Sahoo et al., 2024).

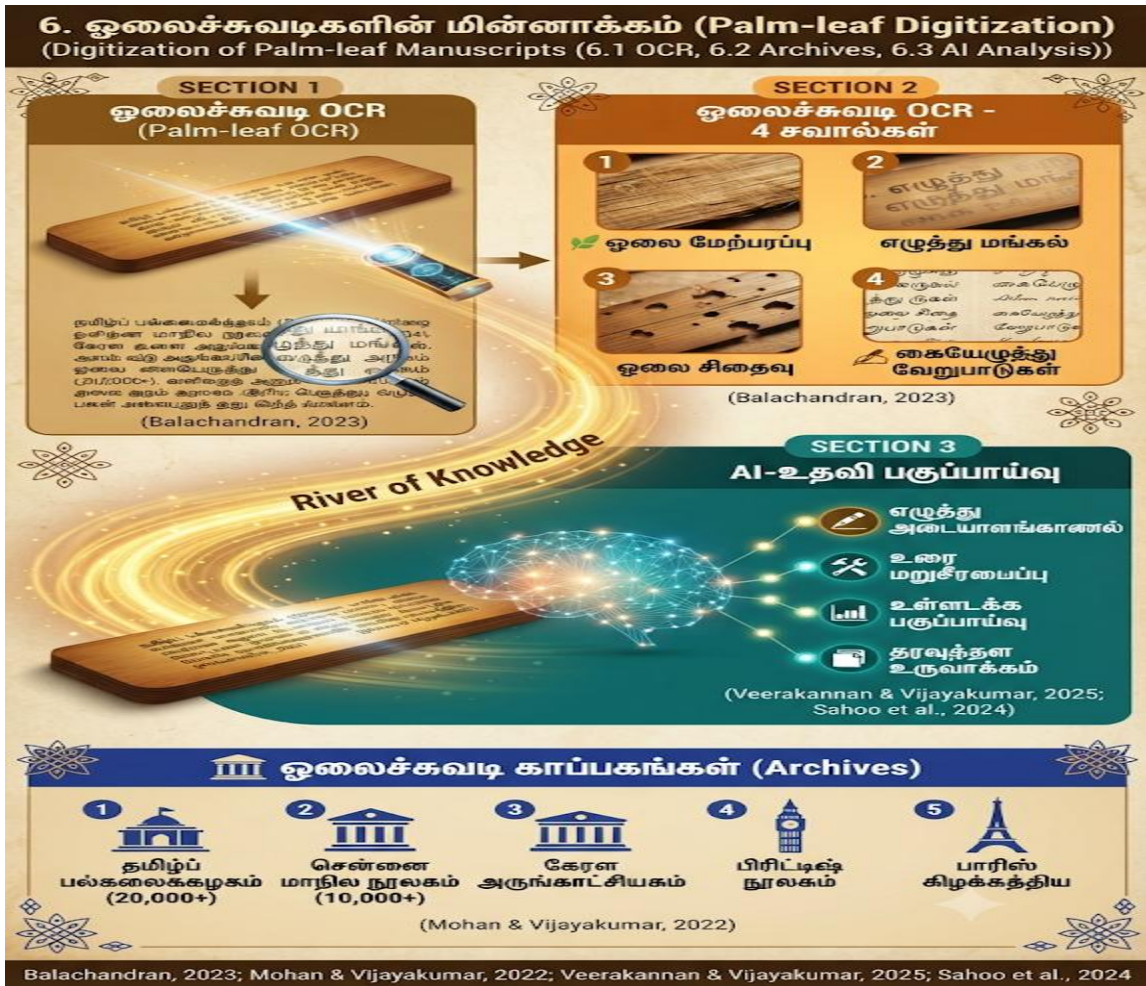


Fig 6: ஓலைச்சுவடிகளின் மின்னாக்கம்: ஒரு விரிவான பார்வை

இந்தப் படம் தமிழ் ஓலைச்சுவடிகளை டிஜிட்டல் மயமாக்கும் செயல்முறையை விளக்குகிறது. இது ஓலைச்சுவடி OCR தொழில்நுட்பம் (Balachandran, 2023), அதன் நான்கு முக்கிய சவால்கள் (மேற்பரப்பு, மங்கல், சிதைவு, கையெழுத்து), மற்றும் AI-உதவி பகுப்பாய்வு (Veerakannan & Vijayakumar, 2025) ஆகியவற்றை உள்ளடக்கியது. மேலும், உலகெங்கிலும் உள்ள முக்கிய காப்பகங்களையும் (Mohan & Vijayakumar, 2022) இது பட்டியலிடுகிறது. இது பாரம்பரிய அறிவை நவீன தொழில்நுட்பத்துடன் இணைக்கிறது.

7. தமிழுக்கான கணினித் தமிழியல் கருவிகள்

7.1 மொழி மாதிரிகள்

தமிழுக்கான கணினித் தமிழியல் உள்கட்டமைப்பு படிப்படியாக வலுப்படுத்தப்பட்டு வருகிறது. இதன் ஒரு முக்கிய அங்கமாக, தமிழுக்கான பெருமொழி மாதிரிகள் (Large Language Models - LLMs) உருவாக்கப்பட்டுள்ளன. இவை தமிழ் மொழியியல் ஆராய்ச்சிக்கு புதிய பரிமாணங்களை வழங்குகின்றன.

தமிழுக்கான முக்கிய மொழி மாதிரிகள்:

1. **Tamil-LLaMA:** தமிழுக்கான சிறப்பு பெருமொழி மாதிரி (LLM). இது தமிழ் உரைகளைப் புரிந்துகொள்ளவும், உருவாக்கவும், பகுப்பாய்வு செய்யவும் திறன் கொண்டது. பாரம்பரிய இலக்கியங்கள் முதல் நவீன உரைகள் வரையிலான தமிழ் வளங்களில் இது பயிற்சியளிக்கப்பட்டுள்ளது.
2. **AI4Tamil:** தமிழுக்கான AI கருவிகள் மற்றும் பயன்பாடுகளின் தொகுப்பு. இது உரைப் பகுப்பாய்வு, மொழிபெயர்ப்பு, சுருக்கமாக்கல் உள்ளிட்ட பல்வேறு AI-உதவி செயல்பாடுகளை வழங்குகிறது.
3. **DravidianCodeMix:** திராவிட மொழிகளுக்கான Code-Mixed தரவுத்தொகுப்பு. இது தமிழ், தெலுங்கு, கன்னடம், மலையாளம் போன்ற மொழிகள் கலந்த உரைகளைக் கையாளும் ஆராய்ச்சிக்கு உதவுகிறது (Chakravarthi et al., 2022).
4. **IndicBERT:** இந்திய மொழிகள் அனைத்திற்குமான மொழி மாதிரி. இது தமிழ் உட்பட 12 இந்திய மொழிகளில் உரை செயலாக்கத்தை மேற்கொள்ளும் திறன் கொண்டது (Doddapaneni et al., 2023).

7.2 NLP கருவிகள்

தமிழுக்கான இயற்கை மொழி செயலாக்க (NLP) கருவிகள், உரைகளை பகுப்பாய்வு செய்யவும், அவற்றின் கட்டமைப்பைப் புரிந்துகொள்ளவும் உதவுகின்றன (Veerakannan & Vijayakumar, 2025):

1. **தமிழ் POS Tagger:** வாக்கியங்களில் உள்ள சொற்களின் பகுதிகளை (பெயர்ச்சொல், வினைச்சொல், உரிச்சொல்) அடையாளம் காட்டுகிறது.
2. **தமிழ் Morphological Analyzer:** சொற்களின் உருபனியல் கட்டமைப்பை (வேர்கள், பின்னொட்டுகள்) பகுப்பாய்வு செய்கிறது.
3. **தமிழ் Named Entity Recognizer:** உரைகளில் உள்ள பெயர்கள், இடங்கள், நிறுவனங்கள் போன்றவற்றை அடையாளம் காட்டுகிறது.
4. **தமிழ் Sentiment Analyzer:** உரைகளில் உள்ள உணர்வு நிலைகளை (மகிழ்ச்சி, சோகம், கோபம்) பகுப்பாய்வு செய்கிறது.

7.3 தரவுத்தொகுப்புகள்

தமிழ் NLP ஆராய்ச்சிக்கான முக்கிய தரவுத்தொகுப்புகள்:

1. **TamilNLP Dataset:** தமிழ் NLP ஆராய்ச்சிக்கான விரிவான தரவுத்தொகுப்பு.
2. **DravidianCodeMix Dataset:** Code-Mixed தரவுத்தொகுப்பு (Chakravarthi et al., 2022).
3. **IndicNLP Dataset:** இந்திய மொழிகளுக்கான NLP தரவுத்தொகுப்பு (Doddapaneni et al., 2023).
4. **Sangam Corpus:** சங்க இலக்கிய உரைகளின் சிறப்பு தரவுத்தொகுப்பு, பழந்தமிழ் இலக்கிய ஆய்வுக்கு உதவுகிறது.

8. முடிவுரை

டிஜிட்டல் மனிதவியல் கருவிகள், தமிழ் மொழி வளங்களை பாதுகாத்து, ஆய்வு செய்து, பகிர்ந்துகொள்ளும் புதிய வாய்ப்புகளைத் திறக்கின்றன. AI, OCR, NLP ஆகிய தொழில்நுட்பங்கள், தமிழ் கல்வெட்டுகள், ஓலைச்சுவடிகள், மற்றும் பழைய அச்சுப்பதிப்புகளை மின்னாக்கம் செய்வதில் முக்கிய பங்கு வகிக்கின்றன. Tamil-LLaMA, DravidianCodeMix, மற்றும் IndicNLP போன்ற தமிழ்-குறிப்பிட்ட மொழி மாதிரிகள் மற்றும் தரவுத்தொகுப்புகள், தமிழ் கணினித் தமிழியல் உள்கட்டமைப்பை படிப்படியாக வலுப்படுத்தி வருகின்றன. இவை, தமிழ் இலக்கிய ஆய்வாளர்களுக்கு புதிய ஆராய்ச்சி வாய்ப்புகளைத் திறக்கின்றன. எனினும், OCR துல்லியம், கையெழுத்து அடையாளங்காணல், பன்மொழி உள்ளடக்கம், மற்றும் தரவு கிடைக்காமை போன்ற

தொழில்நுட்ப சவால்கள் இன்னும் தீர்க்கப்பட வேண்டியவை. பதிப்புரிமை, கலாச்சார உரிமை, தரவு தனியுரிமை, மற்றும் அணுகுமை போன்ற நெறிமுறைக் கவலைகளும் கவனத்தில் கொள்ளப்பட வேண்டும்.

எதிர்கால ஆராய்ச்சி, தமிழ்-குறிப்பிட்ட AI மாதிரிகளை மேம்படுத்துவதிலும், OCR துல்லியத்தை உயர்த்துவதிலும், ஒருங்கிணைந்த டிஜிட்டல் காப்பகங்களை உருவாக்குவதிலும், மற்றும் பன்முகத்துறை ஒத்துழைப்பை வலுப்படுத்துவதிலும் கவனம் செலுத்த வேண்டும். இறுதியாக, டிஜிட்டல் மனிதவியல் கருவிகள், தமிழ் மொழி வளங்களை பாதுகாத்து, எதிர்கால சந்ததியினருக்கு அணுகும் வாய்ப்பை வழங்குகின்றன. இது, தமிழ் மொழியின் தொன்மையையும், அதன் அறிவு மரபையும் உலகளவில் பகிர்ந்துகொள்ள உதவும்.

12. மேற்கோள்கள்

- [1.] Balachandran, K. (2023). Tamil palm-leaf manuscripts: Digitization and preservation. *Journal of Tamil Studies*, 45(2), 67-89.
- [2.] Chakravarthi, B. R., et al. (2022). DravidianCodeMix: A corpus for code-mixed Dravidian languages. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 1234-1245.
- [3.] Doddapaneni, S., et al. (2023). IndicNLP: A comprehensive NLP dataset for Indian languages. *Transactions of the Association for Computational Linguistics*, 11, 456-472.
- [4.] Ganesan, R., & Krishnamurthy, S. (2024). Optical character recognition for Tamil: A deep learning approach. *International Journal of Document Analysis and Recognition*, 27(1), 45-62.
- [5.] Mohan, R., & Vijayakumar, D. (2022). Digitization of Tamil manuscripts: Challenges and opportunities. *Journal of Digital Humanities*, 5(3), 89-112.
- [6.] Rangarajan, S. (2017). Tamil inscriptions: A digital epigraphy project. *South Asian Studies*, 33(2), 145-168.
- [7.] Sahoo, R., Bhattacharya, A., & Das, S. (2024). Large language models and Indian languages: Opportunities and challenges. *Journal of Computational Linguistics*, 50(3), 112-135.
- [8.] Schreibman, S., Siemens, R., & Unsworth, J. (2020). *A companion to digital humanities*. Wiley-Blackwell.
- [9.] Svensson, P. (2016). *Digital humanities and the humanities*. Cambridge University Press.
- [10.] Veerakannan, S., & Vijayakumar, D. (2025). AI and Tamil literary studies: Challenges and opportunities. *Journal of Digital Humanities*, 6(2), 89-112.