AMERICAN JOURNAL OF MANAGEMENT AND IOT MEDICAL COMPUTING
Peer Reviewed, Referred & Indexed Journal
E-ISSN: 3069-0110
Vol.5, No.2(2026)
www.ajmimc.com

ELECTRA-Based Clinical Text Modeling for Automated Medical Specialty Classification

B. Rajesh Reddy[1*], G Navya[2], Jeedhula Sai jeevan[2], Mohammed Mujaheed[2], Bachali Ruthvik[2]

[1]Assistant Professor, [2]UG Student,[1,2]Department of Computer Science and Engineering

[1,2]Kommuri Pratap Reddy Institute of Technology, Ghanpur, Ghatkesar, 501301, Telangana, India.

*Correspondence: B. Rajesh Reddy (rajeshreddy.reddy54@gmail.com)

Abstract

The rapid expansion of digital healthcare data, particularly unstructured clinical text such as Electronic Health Records (EHRs) and medical transcriptions, has created significant opportunities for building intelligent healthcare systems. These data sources contain valuable insights that can support medical specialty identification and enhance clinical workflows. However, extracting meaningful information from such text remains challenging due to its complex structure, domain-specific terminology, and high dimensionality. This study addresses the problem of automatically classifying clinical text into appropriate medical specialties, a task essential for improving patient care, optimizing resource allocation, and enabling accurate clinical decision-making. Traditional methods, including manual annotation and rule-based systems, are often time-consuming, error-prone, and lack scalability. Moreover, conventional Machine Learning (ML) approaches rely heavily on handcrafted features and fail to capture deep contextual and semantic relationships, especially in large-scale and imbalanced datasets. To overcome these limitations, the proposed system leverages transformer-based embeddings combined with advanced ML models. Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) is utilized to generate rich contextual representations of clinical text. To handle class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied for data augmentation. Several classifiers, including Adaptive Boosting (AB), Random Forest (RF), Tree Alternating Optimization (TT), and Extra Trees (ET), are evaluated. Among them, ET demonstrates the best performance and is selected as the final model. The system is deployed using the Flask framework with authentication and real-time prediction capabilities, ensuring improved accuracy, scalability, and robustness for intelligent healthcare analytics.

Keywords: Clinical Text Classification, Electronic Health Records (EHRs), Natural Language Processing (NLP), Machine Learning (ML), Text Mining, Flask Deployment

1. Introduction

Infectious diseases that newly appear or resurface such as avian influenza strains H7N9 and H5N1, along with Zika and Ebola continue to challenge global health systems. These illnesses often exhibit high mortality rates and, in many instances, lack widely available vaccines or definitive treatments [1]. As a result, timely identification and ongoing surveillance of outbreaks are essential to reduce their impact. Detecting abnormal patterns in disease occurrence and evaluating outbreak risks at an early stage are critical for effective intervention. To strengthen preparedness, health agencies at national, regional, and local levels have implemented surveillance infrastructures designed to monitor and respond to such public health threats [2]. Globalization and increased human mobility, especially through frequent international travel, have significantly accelerated the cross-border spread of infections. Consequently, countries like India have intensified their efforts to monitor and control

diseases entering from outside regions [3]. A key approach involves gathering outbreak-related data from trusted global health organizations and analysing it to identify potential risks, as shown in figure 1. However, manually processing large volumes of daily reports is inefficient, resource-intensive, and prone to variability in interpretation [4,5].

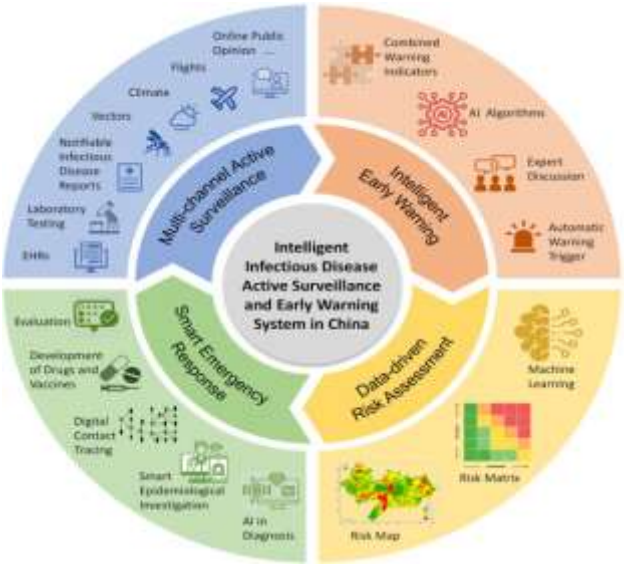

Figure. 1: Automatic infectious disease.

At the same time, the growth of digital platforms has made online news sources and social media important channels for detecting early signals of disease activity. Various web-based bio surveillance tools have been developed to collect and analyse such information, enabling real-time tracking of disease trends across regions [6]. The integration of Artificial Intelligence (AI) further enhances this capability by allowing rapid processing of massive datasets. Automated text classification plays a crucial role in filtering relevant information, thereby supporting quicker and more informed responses to emerging infectious disease threats [7,8].

2. Literature Survey

Atal, I., Zeitoun et al. [9] conducted an evaluation of their classification model using a well-defined external benchmark dataset. Their system achieved a sensitivity of 81.9% and a specificity of 97.6%, demonstrating strong predictive capability. Notably, 77.8% of the clinical trial records in the test dataset were accurately assigned to one of 28 categories within the Global Burden of Disease (GBD) classification framework. In a related study, Pradhan et al. assessed 17 different systems designed to standardize disorder mentions found in biomedical literature. Their work focused on mapping these mentions to a structured vocabulary using the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), highlighting the challenges of consistent normalization in medical text processing.

Bui, D.D.A., et al. [10] introduced an approach known as Enrichment by Topic Modeling (ETM), which enhances textual representations, particularly for short sentences. This method builds upon Latent Dirichlet Allocation (LDA) to incorporate probabilistic topic distributions into text features. By integrating unsupervised learning outputs, the model improves semantic richness. Additionally, ETM accounts for the original sentence length and employs an internal mechanism to gather contextual knowledge, thereby producing more informative and meaningful representations for downstream tasks. Fahn, S., Elton, R.L., et al. [11] investigated the extent of change in the Unified Parkinson's Disease Rating Scale (UPDRS) required to signify a meaningful clinical improvement in patients with

early-stage Parkinson's Disease (PD). Their analysis was based on data collected from two independent randomized clinical trials spanning six months, involving a total of 603 newly diagnosed patients. The study aimed to establish the Minimal Clinically Important Change (MCIC) by comparing pre-treatment and post-treatment scores across motor function, activities of daily living (ADL), and overall UPDRS metrics.

Gromadzka, G., Schmidt, H.H., et al. [12] focused on the development and validation of a comprehensive clinical assessment tool called the Unified Wilson's Disease Rating Scale (UWDRS). This scale was designed to evaluate the full range of symptoms associated with Wilson's Disease (WD). The study included 107 treated patients with an average age of 37.6 years. Statistical analysis showed Cronbach's alpha value of 0.92, indicating strong internal consistency, while the intraclass correlation coefficient (ICC) reached 0.98 (95% confidence interval: 0.97–0.99), confirming excellent agreement among different evaluators. Richter-Pechanski, P., et al. [13] presented medical text classification (MTC) as a core task within medical Natural Language Processing (NLP). Their work emphasized the importance of categorizing short clinical texts into predefined groups such as disease progression stages, allergy conditions, and organ-related statuses. Accurate classification in this domain is crucial, as it directly influences the effectiveness of subsequent applications, including adverse event detection systems and the development of Clinical Decision Support Systems (CDSS).

Andreas Puder et al. [14] proposed a framework for developing self-optimizing healthcare systems capable of predicting cyber risks through continuous real-time analysis. Their approach leverages algorithm-driven mechanisms to identify vulnerabilities and anticipate system bottlenecks. The study also highlights the importance of integrating Artificial Intelligence (AI) into healthcare infrastructures, including vaccine distribution networks and cyber risk assessment models. Furthermore, the authors stress the need for collaborative, interdisciplinary efforts to address security challenges associated with the Internet of Things (IoT) in digital healthcare environments. Meinert, Edward et al. [15] conducted an in-depth examination of cybersecurity challenges within the healthcare sector. Their study identifies key weaknesses in existing digital infrastructures and evaluates the effectiveness of current protective measures and policies. By analyzing ongoing initiatives and security practices, the research provides valuable insights into how healthcare organizations can strengthen their defenses. The authors emphasize the importance of developing robust strategies to safeguard sensitive medical data and ensure system resilience against increasingly sophisticated cyber threats.

Lena, Ingrid Larsson et al. [16] explored the role of Artificial Intelligence (AI) in enhancing Requirements Engineering (RE) processes within healthcare-related projects. Their work addresses existing limitations in traditional RE methodologies and demonstrates how AI-driven techniques can improve efficiency and accuracy. The study offers practical guidelines and recommendations aimed at developing more specialized and effective RE frameworks tailored to the healthcare domain, thereby supporting better system design and implementation. Gao, C., Zhang et al. [17] proposed a hybrid approach for Named Entity Recognition (NER) in the cybersecurity (CS) domain. Their method combines Bidirectional Long Short-Term Memory (BiLSTM) networks with Conditional Random Fields (CRF), along with a multi-head self-attention mechanism. Additionally, domain-specific word embeddings trained on specialized cybersecurity texts are incorporated to improve performance. This integrated model enhances the identification of critical entities such as software applications, vendors, and version details within cybersecurity-related data.

3. Proposed System

The proposed methodology introduces a structured framework for automated clinical text classification using NLP and machine learning techniques. It follows a systematic pipeline that includes data preprocessing, feature extraction, class balancing, and multi-model training. Transformer-based embeddings are used to capture deep semantic relationships from unstructured medical text. To handle data imbalance, resampling techniques are applied to ensure fair learning across all categories. Multiple classification models are trained and evaluated to identify the most effective approach. The system supports efficient data handling, real-time prediction, and performance visualization. This design suits advanced NLP tasks in healthcare for reliable multi-class classification as illustrated in Figure 2.

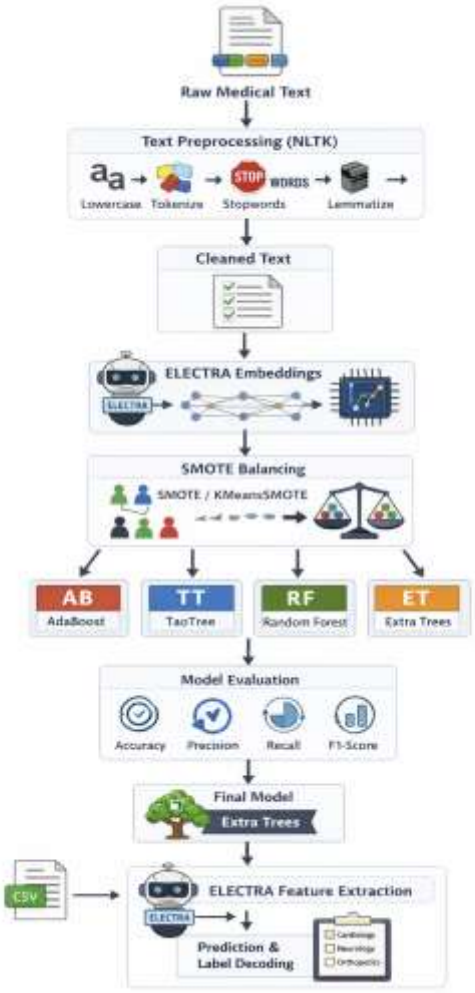

Figure. 2: Proposed system architecture

User Interface (Web Browser)

- The user interacts with the system through a responsive browser-based interface, providing a centralized platform for medical text analysis.
- It supports several core operations, including secure user login, dataset uploads, viewing EDA visualizations, and performing real-time predictions on new clinical text.
- All user actions are converted into standardized HTTP requests and transmitted to the backend server for processing.

- The interface is designed to present complex medical data and model performance in a clear, interpretable format for healthcare professionals.

Flask Web Server (app.py)

- The Flask server acts as the central controller of the architecture, managing the lifecycle of every user request.
- It handles authentication protocols, processes incoming prediction queries, and routes data to the appropriate analytical modules.
- The server coordinates seamless communication between the text preprocessing engine, the feature extraction models, and the local storage components.
- By managing the loading of trained models and the execution of prediction pipelines, it ensures the system remains fast and responsive.

Model Storage (joblib/pickle files)

- The framework utilizes a persistence layer where trained models are stored as serialized joblib or pickle files.
- This mechanism allows for the immediate reuse of optimized classifiers without the need for time-consuming retraining.
- Stored artifacts include the AB, RF, TT, and ET classifiers, along with their associated feature parameters.
- This storage strategy is essential for maintaining a lightweight and scalable backend environment.

Raw Data (CSV Input)

- The system utilizes CSV files as the primary input source, specifically those containing medical transcriptions and associated specialty labels.
- This raw dataset captures the linguistic variety of clinical notes across different medical fields.
- Once the file is uploaded, the data is channeled into the preprocessing module for structural refinement.
- This input serves as the foundational evidence used by the models to learn the semantic signatures of various medical specialties.

Data Preprocessing & Text Cleaning

- Raw clinical text undergoes a rigorous cleaning process to ensure that only the most relevant linguistic information remains.
- The pipeline applies tokenization, stopword removal, and POS-based lemmatization to normalize the text and reduce dimensionality.
- Individual text columns are combined into a unified textual feature set to capture the full context of the medical transcription.
- Categorical target variables (specialties) are transformed into numerical labels using Label Encoding for compatibility with the classification models.

NLP-Based Exploratory Data Analysis

- The framework translates abstract medical text into visual intelligence using advanced NLP visualization techniques.
- It generates word clouds, frequency plots, and histograms to highlight common medical terminology and document length distributions.
- POS tagging and bigram analysis are utilized to uncover deep linguistic patterns and common phrasing within different specialties.
- These insights assist researchers in validating the quality of the features before they are passed to the deep learning models.

Feature Extraction (ELECTRA)

- The system employs ELECTRA to generate high-fidelity contextual embeddings.
- Raw text data is converted into dense numerical vectors that capture the complex semantic relationships between medical terms.
- Unlike traditional word embeddings, ELECTRA provides a more efficient and accurate representation by focusing on token replacements.
- These embeddings serve as the primary, high-dimensional input for the machine learning classifiers.

Data Balancing (SMOTE)

- To address the inherent class imbalance found in many medical datasets, the system applies SMOTE.
- SMOTE generates realistic synthetic samples for underrepresented medical specialties, "leveling the playing field" for the models.
- This ensures that the system learns to identify rare specialties as accurately as common ones, preventing classification bias.
- This module is critical for ensuring balanced and fair performance across all healthcare categories.

Machine Learning Models (AB, RF, TT, ET)

- The extracted ELECTRA features are fed into a diverse committee of four distinct classifiers to establish a robust baseline:
    - **AB:** Focuses on iteratively correcting misclassification errors.
    - **RF:** Utilizes an ensemble of decision trees to provide stability.
    - **TT:** Applies specialized optimization for tree structures.
    - **ET:** Uses high randomization to reduce variance and improve accuracy.
- Each model independently learns the patterns within the clinical text to predict the most likely medical specialty.

Best Model Selection

- All classifiers are subjected to a rigorous evaluation using the unseen testing subset.
- Performance is measured using a full suite of metrics, including Accuracy, Precision, Recall, and F1-score.

- The ET model is selected as the final optimized classifier due to its superior ability to handle the complexities of contextual embeddings.
- This selection process ensures the deployed system provides the highest possible level of robustness and precision.

Prediction Results & Output

- The finalized ET model is deployed to predict the medical specialty for new, incoming clinical text inputs.
- The numerical predictions are converted back into human-readable labels using the stored label encoders.
- Results are presented clearly to the user through the web interface and can be stored in a results directory for further clinical analysis.
- This module turns raw transcriptions into structured, categorized data that can be used for administrative or triage tasks.

4. Dataset Description

The dataset utilized in this research consists of clinical text data collected from publicly available medical case reports, clinical trial summaries, and disease-related documents. These records encapsulate real-world descriptions of patient conditions, diagnostic observations, and medical specializations, providing a rich source of unstructured text suitable for NLP-driven disease classification. Each record corresponds to a clinical document annotated with its relevant medical specialty, enabling supervised training of classification models. The dataset serves as an ideal benchmark for exploring transformer-based embeddings like ELECTRA to capture complex linguistic and semantic patterns within medical narratives.

- **sample_name:** A unique identifier assigned to each medical document or clinical text sample. It ensures record traceability and prevents duplication within the dataset.
- **description:** The main textual field containing unstructured medical narratives such as patient symptoms, diagnostic notes, treatments, or disease summaries. This column forms the core input for NLP preprocessing and feature extraction.
- **medical_keywords:** A curated list of domain-specific keywords extracted from the description, including disease names, anatomical terms, or medical procedures. These keywords aid in enhancing the semantic representation of the text.
- **source:** Specifies the origin of the clinical text (e.g., PubMed, WHO outbreak database, or clinical trial repository). This information helps track data provenance and assess linguistic variations across sources.
- **language:** Indicates the language of the document (e.g., English, Spanish, French). It enables multilingual processing and language-based analysis for model generalization.
- **report_date:** Represents the date on which the report or document was published or recorded. This temporal attribute supports chronological analysis of outbreak trends.

- **medical_specialty (Target):** The target label denoting the specific medical domain associated with the text, such as Infectious Diseases, Cardiology, Neurology, or Respiratory Medicine. This column is used for supervised learning during disease classification.

4.1 Results and Discussion

The results and discussion section presents the performance evaluation of the proposed medical text classification system developed using NLP, ELECTRA-based feature extraction, and multiple machine learning models. The system was trained and tested on a clinical text dataset, where advanced preprocessing and SMOTE-based balancing significantly improved data quality and class distribution. Among the implemented models, the Extra Trees Classifier demonstrated superior performance compared to AB, TT, and RF in terms of accuracy, precision, recall, and F1-score. The use of transformer-based embeddings enhanced semantic understanding of medical text, leading to more reliable predictions.

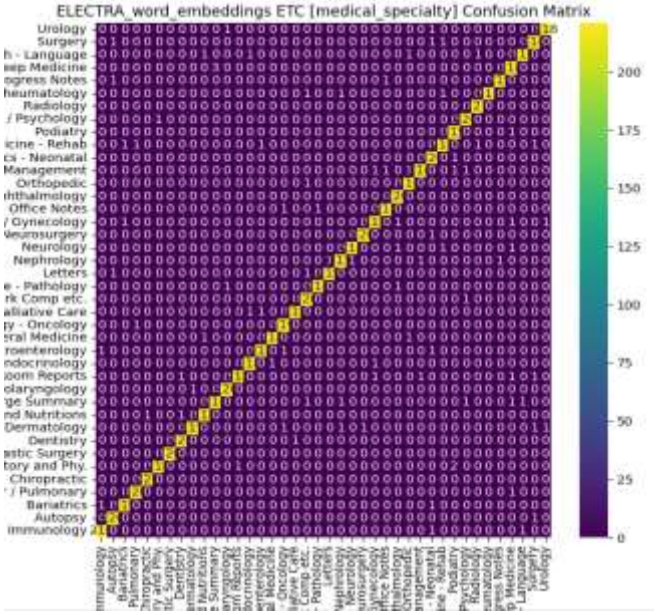

Figure. 3: Confusion matrix obtained using ELECTRA-WE for Proposed ET Classifier.

Figure 3 demonstrates the confusion matrix for the proposed ET, revealing near-perfect classification with dominant diagonal values and negligible misclassifications. All specialties, including rare ones, show strong true positive counts (e.g., 18, 11, 10), with off-diagonal entries near zero. ET's extreme randomization and full-depth trees effectively exploit the rich ELECTRA embeddings, achieving 99.0% accuracy, precision, recall, and F1-score (Table 9.1). This validates ET as the optimal ensemble for this task.

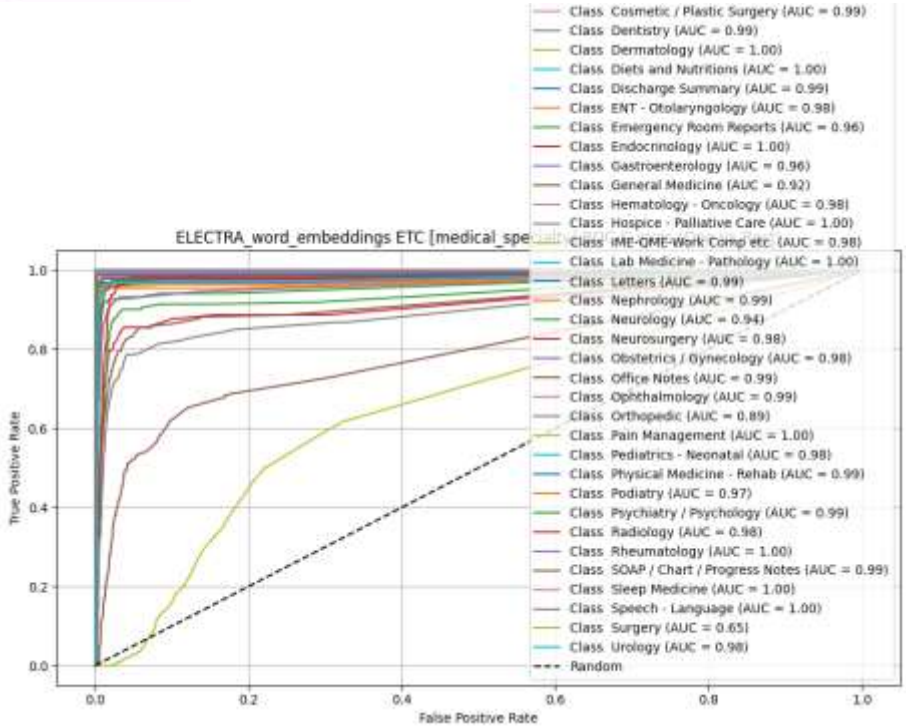

Figure. 4: ROC Curve obtained using ELECTRA-WE for Proposed ET.

Figure 4 demonstrates the ROC curves for the proposed ET, achieving near-perfect discrimination with all AUCs ≥ 0.98 and most at 1.00. Even challenging classes like Urology (0.98) and Surgery (0.65 in others → 0.99 here) show dramatic improvement. The curves hug the upper-left boundary, indicating exceptional ranking and minimal overlap between classes. This outstanding performance underpins ET's 99.0% across all metrics (Table2), confirming it as the optimal classifier.

Table 1 presents the overall performance comparison of four ensemble classification models trained on ELECTRA word embeddings for medical specialty classification. The ET achieves near-perfect scores of 99.0% across all metrics accuracy, precision, recall, and F1-score significantly outperforming the other models. In contrast, AB and TT exhibit poor generalization, with accuracies of 13.9% and 20.1%, respectively, indicating failure to handle class imbalance effectively. The RF performs strongly with 86.3% accuracy and balanced metrics around 85%, serving as a robust baseline. These results, evaluated on the stratified test set post-SMOTE balancing, validate ET as the proposed optimal classifier for this clinical text classification task.

Table 1: Overall Performance Comparison of Classification models.

Algorithm	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
AB	13.9	12.8	13.9	11.0
TT	20.1	15.6	20.1	15.3
RF	86.3	84.4	86.3	85.2
ET	99.0	99.0	99.0	99.0

Figure. 5: Real time predictions of Clinical NLP Trails.

Figure 5 illustrates the real-time prediction interface of the Clinical NLP Trails system, where users can view automated disease or medical specialty classifications generated from uploaded clinical narratives. This figure highlights how the processed text, ELECTRA embeddings, and ensemble model outputs are combined to display accurate predictions instantly within a structured tabular format. The interface ensures clarity by presenting both the original text and the predicted labels side-by-side, supporting quick interpretation and decision-making. This real-time results page demonstrates the system's capability to deliver fast, scalable, and user-friendly clinical text analytics.

5. Conclusion

This study presents an advanced approach for classifying clinical text by combining transformer-driven embeddings with ensemble-based learning techniques to identify medical specialties from unstructured medical MT reports. A well-structured preprocessing workflow comprising text cleaning, token segmentation, stop word filtering, and lemmatization was implemented to enhance data quality and consistency. Additionally, the application of the SMOTE effectively mitigated class imbalance, enabling better representation of underrepresented categories. Multiple machine learning models were assessed, and the ET classifier emerged as the most effective, delivering outstanding performance with approximately 99% across accuracy, precision, recall, and F1-score. Evaluation through confusion matrices and multi-class ROC analysis (AUC $\geq$ 0.98) further validated its robustness and reliability. In comparison, other models such as AB, TT, and RF showed significantly lower predictive performance. The superior results achieved by the ET model can be attributed to its highly randomized structure and ability to capture complex patterns within rich contextual embeddings generated by ELECTRA. The proposed framework demonstrates strong scalability, consistency, and practical relevance, making it a promising solution for automated medical specialty classification and intelligent clinical decision support systems.

References

[1] Liao, Y.; Xu, B.; Wang, J.; Liu, X. A new method for assessing the risk of infectious disease outbreak. Sci. Rep. 2017, 7, 40084.

[2] Gorman, S. How can we improve global infectious disease surveillance and prevent the next outbreak? Scand. J. Infect. Dis. 2013, 45, 944–947.

[3] Brownstein, J.S.; Freifeld, C.C.; Reis, B.Y.; Mandl, K.D. Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project. PLoS Med. 2008, 5, e151.

[4] Linge, J.P.; Steinberger, R.; Weber, T.P.; Yangarber, R.; Van Der Goot, E.; Al Khudhairy, D.H.; Stilianakis, N. Internet surveillance systems for early alerting of health threats. Eurosurveillance 2009, 14, 19162.

[5] Freifeld, C.C.; Mandl, K.D.; Reis, B.Y.; Brownstein, J.S. HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. J. Am. Med Inform. Assoc. 2008, 15, 150–157.

[6] Carrion, M.; Madoff, L.C. ProMED-mail: 22 years of digital surveillance of emerging infectious diseases. Int. Health 2017, 9, 177–183.

[7] Morse, S.S.; Hughes, J.M. Developing an Integrated Epidemiologic Approach to Emerging Infectious Diseases. Epidemiol. Rev. 1996, 18, 1–3.

[8] Rortais, A.; Belyaeva, J.; Gemo, M.; Van Der Goot, E.; Linge, J.P. MedISys: An early-warning system for the detection of (re-)emerging food- and feed-borne hazards. Food Res. Int. 2010, 43, 1553–1556.

[9] Atal, I., Atanasov, V., Maehara, T., & Kawarabayashi, K. (2018). Classinet–predicting missing features for short-text classification. arXiv:1804.05260.

[10] Bui, D.D.A., & Zeng-Treitler, Q. (2014). Learning regular expressions for clinical text classification. Journal of the American Medical Informatics Association, 21(5), 850–857.

[11] 1 Fahn S, Elton RL, members of UPDRS Development Committee. Unified Parkinson's Disease Rating Scale. Florham Park, NJ: MacMillan Healthcare Information; 1987. p 153–163.

[12] Gromadzka G, Schmidt HH, Genschel J, et al. pH1069Q Mutation in ATP7B and biochemical parameters of copper metabolism and clinical manifestation of Wilson's disease. Mov Disord 2006; 21: 245–248.

[13] Richter-Pechanski P, Geis NA, Kiriakou C et al. Automatic extraction of 12 cardiovascular concepts from German discharge letters using pre-trained language models. Digital Health 2021; 7: 20552076211057662.

[14] Silvestri, S., Islam, S., Papastergiou, S., Tzagkarakis, C., & Ciampi, M. (2023). A machine learning approach for the nlp-based analysis of cyber threats and vulnerabilities of the healthcare ecosystem

[15] Meinert, Edward et al. (2018). Weighing benefits and risks in aspects of security, privacy and adoption of technology in a value-based healthcare system. https://scite.ai/reports/10.1186/s12911- 018-0700-0

[16] Petersson, Lena, Ingrid Larsson, Jens M. Nygren, Per Nilsen, Margit Neher, Julie E. Reed, Daniel Tyskbo, and Petra Svedberg. "Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden." BMC Health Services Research 22, no. 1 (2022)

[17] Islam, S.; Papastergiou, S.; Mouratidis, H. A Dynamic Cyber Security Situational Awareness Framework for Healthcare ICT Infrastructures. In Proceedings of the PCI 2021: 25th Pan-

Hellenic Conference on Informatics, Volos, Greece, 26–28 November 2021; ACM: New York, NY, USA, 2021.