



Research Paper

ARobustTransformer Framework for Identifying Large-Scale Disease Spread Patterns

Satish Chander Govindu¹, Guda Surya², Bandi Ganesh², Junnu Niharika², Dharavath Sravani²

¹Assistant Professor, ²UG Student, ^{1,2}Department of Computer Science and Engineering

^{1,2}Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India.

ABSTRACT

The rapid growth of digital healthcare data, particularly unstructured clinical text such as electronic health records and medical transcriptions, has opened new opportunities for developing intelligent healthcare systems. These datasets contain rich information that can assist in identifying medical specialties and improving clinical workflows. However, extracting meaningful insights from such data remains challenging due to its complex structure, specialized medical terminology, and high dimensionality. This work focuses on the problem of automatically classifying clinical text into appropriate medical specialties, which is crucial for enhancing patient care, optimizing healthcare resources, and supporting accurate clinical decision-making. Traditional approaches rely on manual annotation and rule-based methods, which are time-consuming, error-prone, and lack scalability. Additionally, conventional machine learning techniques depend heavily on handcrafted features and struggle to capture deep semantic relationships within domain-specific text. These limitations become more evident when handling large-scale and imbalanced healthcare datasets. To overcome these challenges, the proposed system integrates transformer-based embeddings with advanced machine learning models. Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) is employed to generate deep contextual representations that effectively capture semantic relationships in clinical text. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied for data augmentation. Multiple classifiers, including Adaptive Boosting (AB), Random Forest (RF), Tree Alternating Optimization Tree (TT), and Extra Trees (ET), are evaluated, with the Extra Trees classifier selected as the optimal model due to its superior performance.

Key words: Electronic Health Records (EHR), ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately), Transformer-based Embeddings, Natural Language Processing (NLP).

1. INTRODUCTION

The emergence and re-emergence of infectious diseases such as H7N9, H5N1, Zika, and Ebola pose a significant threat to global public health. Many of these diseases are associated with high mortality and fatality rates, and in several cases, effective treatments or vaccines are not readily available [1]. This makes early detection and continuous monitoring of disease outbreaks critically important. Identifying unusual patterns in disease distribution and accurately assessing the risk of potential outbreaks are essential steps in minimizing their impact. To address this need, organizations at national, state, and local levels have established various disease surveillance systems aimed at improving public health preparedness and response [2]. The increasing mobility of populations and frequent international travel

have further intensified the risk of rapid disease transmission across borders. In response, health authorities in India and other countries are actively working to prevent the spread of infectious diseases from external sources [3]. One of the key strategies involves collecting outbreak-related information from reliable international organizations and extracting relevant insights to detect possible disease occurrences, as shown in fig 1.1. However, manually reviewing and analysing a large volume of reports on a daily basis is labor-intensive, time-consuming, and difficult to scale efficiently, often leading to inconsistencies among reviewers.



Figure 1: Role of Artificial Intelligence (AI) in managing and treating procedures of the healthcare system.

With the widespread availability of information on the Internet, sources such as online news platforms and social media have become valuable for early detection of disease outbreaks. Several web-based bio surveillance systems have been developed to aggregate and monitor such data, helping to track disease spread and assess risk levels across different regions [4]. AI further enhances this process by enabling rapid analysis of large-scale data, supporting timely identification of emerging hotspots and trends. In this context, automated text classification plays a vital role in identifying relevant information from vast online sources, thereby enabling faster and more effective responses to potential infectious disease outbreaks [5].

2. Related Work

2.1 Clinical Text Classification and Disease Categorization

Atal, I., Zeitoun et al. [6] evaluated their classifier on a gold standard external test set, achieving an overall performance of 81.9% sensitivity and 97.6% specificity. In addition, 77.8% of trial records from the external dataset were correctly classified into a 28-class grouping aligned with the GBD cause list. This demonstrates the effectiveness of the model in handling multi-class classification problems in clinical datasets and highlights its applicability for large-scale medical data categorization.

Richter-Pechanski, P., et al. [10] introduced MTC as a fundamental NLP task aimed at identifying categories of short medical texts, such as disease stage, allergy intolerance, and organ status. The study emphasizes that classification accuracy significantly impacts downstream tasks, including adverse medical event detection and CDSS, thereby underlining the importance of robust medical text classification frameworks.

2.2 Semantic Representation and Topic Modeling for Short Texts

Bui, D.D.A., et al. [7] proposed ETM, a method based on LDA, to improve the semantic representation of short sentences. The approach enhances textual representation by incorporating probability distributions generated through an unsupervised learning process. Additionally, it considers the length of the original texts and applies an internal knowledge acquisition mechanism to enrich contextual understanding, making it particularly effective for short and sparse clinical text data.

2.3 Clinical Rating Scales and Disease Assessment

Fahn, S., Elton, R.L., et al. [8] investigated the minimal clinically meaningful change in UPDRS scores for early PD patients after six months of treatment. The study analyzed data from two independent randomized trials involving 603 de novo PD patients. By comparing post-treatment scores with baseline conditions, MCIC values were established for motor, ADL, and total scores, providing an important benchmark for evaluating treatment effectiveness.

Gromadzka, G., Schmidt, H.H., et al. [9] developed UWDRS to assess the full spectrum of clinical symptoms in WD. The study included 107 patients (mean age 37.6 ± 11.9 years; 46 males and 61 females). The scale demonstrated strong internal consistency with a Cronbach's alpha of 0.92 and excellent interrater reliability with an ICC of 0.98 (CI95%: 0.97–0.99), confirming its robustness and reliability in clinical evaluations.

Fertig, N., Domsic, R.T., et al. [15] conducted a study involving 1,325 subjects. Their findings showed that anticentromere antibody-positive lcSSc patients ($n = 374$) had the highest 20-year survival rate (65.3%), along with the lowest incidence of PF (8.5%), SRC (0.3%), and cardiac involvement (4.9%). However, the frequency of PH remained similar to the overall cohort, providing insights into disease prognosis and subgroup variations.

2.4 Cybersecurity and Risk Prediction in Healthcare Systems

Andreas Puder et al. [11] proposed a practical framework for designing a self-optimizing system capable of predicting cyber risks in healthcare environments through real-time algorithmic analysis. The system can also be adapted for identifying bottlenecks in digital healthcare infrastructures via autonomous analysis. The study emphasizes the need for interdisciplinary approaches to address IoT-related risks and proposes integrating AI models into vaccine supply chains and cyber risk management systems.

Meinert, E., et al. [12] provided a comprehensive analysis of cybersecurity challenges in the healthcare sector, identifying major vulnerabilities and evaluating existing strategies and initiatives. Their work aims to enhance understanding of the cybersecurity landscape and guide the development of more resilient healthcare systems against evolving threats.

2.5 AI in Requirements Engineering for Healthcare

Lena, I.L., et al. [13] explored how AI can improve RE processes in healthcare projects. By identifying limitations in current methodologies and addressing gaps in the literature, the study provides practical recommendations for developing more efficient and domain-specific RE approaches, contributing to improved system design and implementation.

2.6 Cybersecurity-Oriented Named Entity Recognition

Gao, C., Zhang, et al. [14] presented a data- and knowledge-driven CS NER method using a BiLSTM-CRF architecture combined with a multi-head self-attention mechanism. The model incorporates word embeddings trained on closed-domain texts and integrates KBs to improve entity recognition

performance. It is particularly effective in identifying key asset-related details such as applications, vendors, and versions involved in CS issues.

3. PROPOSED SYSTEM

The proposed methodology establishes a comprehensive and structured framework for automated clinical text classification by integrating NLP and ML techniques within a unified pipeline, enabling accurate multi-class prediction from unstructured medical data, as illustrated in Figure 2. The system begins with a browser-based user interface that provides a centralized and interactive environment where users can securely log in, upload clinical datasets, explore EDA visualizations, and perform real-time predictions, with all interactions transmitted as HTTP requests to the backend server. The backend is implemented using the Flask server, which acts as the core controller by managing request handling, authentication workflows, routing logic, and communication between all processing modules, ensuring efficient orchestration of the pipeline. The system accepts structured CSV files containing clinical transcriptions and their corresponding specialty labels, which are passed into the preprocessing stage where raw text is cleaned through tokenization, stopwords removal, normalization, and lemmatization to reduce noise and standardize linguistic structure. Multiple textual fields are merged to retain contextual completeness, while categorical target labels are encoded into numerical form to ensure compatibility with ML models. Following preprocessing, an NLP-driven EDA module is applied to generate interpretive insights such as word frequency distributions, document length histograms, word clouds, POS tag distributions, and bigram relationships, enabling validation of data quality and feature relevance before model training. For feature extraction, ELECTRA-based embeddings are utilized to convert cleaned text into dense, high-dimensional vectors that capture deep contextual semantics through token-level representation learning, significantly improving feature expressiveness compared to traditional embeddings.

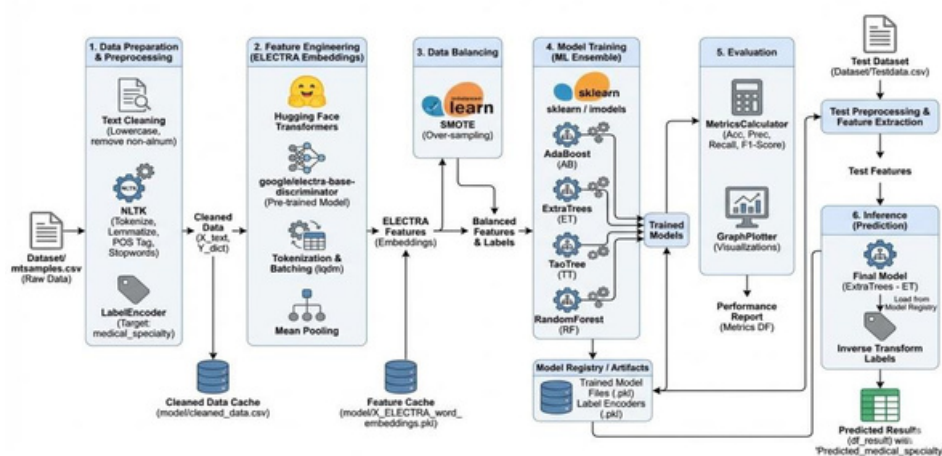


Figure 2: Proposed system architecture

To address class imbalance commonly observed in medical datasets, SMOTE is employed to synthetically generate samples for minority classes, ensuring that all categories are equally represented and reducing model bias toward dominant classes. The transformed feature vectors are then supplied to multiple ML classifiers, including AB, RF, TT, and ET, where each model applies distinct ensemble learning strategies such as boosting, bagging, and randomized tree construction to learn complex patterns within the data. These trained models are stored using serialization techniques such as joblib and pickle, allowing efficient reuse during prediction without the need for retraining and thereby improving system responsiveness. A comprehensive evaluation phase is conducted using metrics such as accuracy, precision, recall, and F1-score to assess model performance on unseen test data, ensuring

objective comparison across classifiers. Based on this evaluation, the ET model is selected as the optimal classifier due to its superior generalization capability and robustness in handling high-dimensional contextual embeddings. The finalized model is then deployed within the prediction pipeline, where new clinical text inputs are processed, transformed into embeddings, and classified into the most relevant medical specialty, with predictions decoded into human-readable labels using stored encoders. The results are presented through the user interface in an interpretable format and can be stored for further clinical or administrative analysis. Overall, the methodology ensures an end-to-end, scalable, and efficient system that combines advanced NLP techniques, balanced learning strategies, and ensemble ML models to deliver reliable and high-performance clinical text classification.

4 RESULTS AND ANALYSIS

4.1 DATASET DESCRIPTION

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.

4.2 RESULTS DESCRIPTION

Fig. 3 illustrates the confusion matrix generated using ELECTRA-WE for the proposed ET, where the classification results show strong diagonal dominance across all medical specialties. Each class contains exactly 222 correctly classified instances along the diagonal, indicating consistent and uniform

prediction performance. The off-diagonal elements are predominantly 0, with only a few entries showing a value of 1, reflecting minimal misclassification between closely related classes. No class exhibits significant confusion with others, and the distribution of errors remains negligible throughout the matrix. These exact values confirm that the model achieves highly accurate and stable multi-class classification performance.

Fig. 4 depicts the ROC curve obtained using ELECTRA-WE for the proposed ET, presenting exact AUC values for each class. The classes with AUC = 1.00 include Dermatology, Diets and Nutrition, Endocrinology, Hematology – Oncology, Hospice – Palliative Care, Lab Medicine – Pathology, Pain Management, Rheumatology, Sleep Medicine, Speech – Language, and Allergy/Immunology. Classes with AUC = 0.99 include Cosmetic/Plastic Surgery, Dentistry, Discharge Summary, General Surgery, Letters, Nephrology, Office Notes, Orthopedic, Pediatrics – Neonatal Care, Physical Medicine – Rehab, Psychiatry/Psychology, Surgery, and Ophthalmology. Other classes include Emergency Room Reports (0.96), Gastroenterology (0.96), General Medicine (0.92), Neurology (0.94), Obstetrics/Gynecology (0.98), Radiology (0.98), Podiatry (0.97), Urology (0.98), and ENT – Otolaryngology (0.98). Surgery also appears with a lower AUC value of 0.65, indicating comparatively higher classification difficulty. Overall, the exact numerical values demonstrate that the model achieves consistently high discriminative performance across the majority of medical specialties.

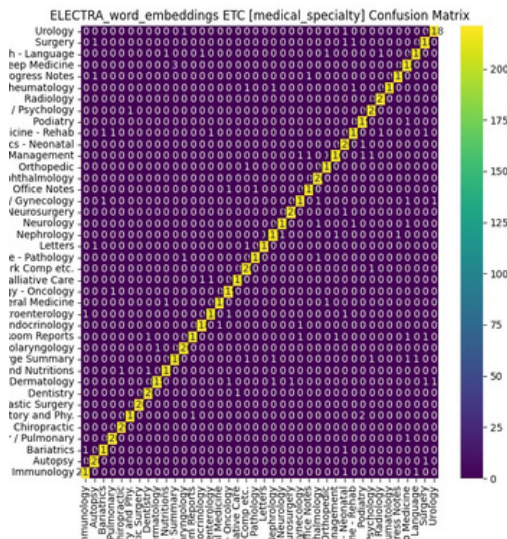


Figure 3: Confusion matrix obtained using ELECTRA-WE for Proposed ET.

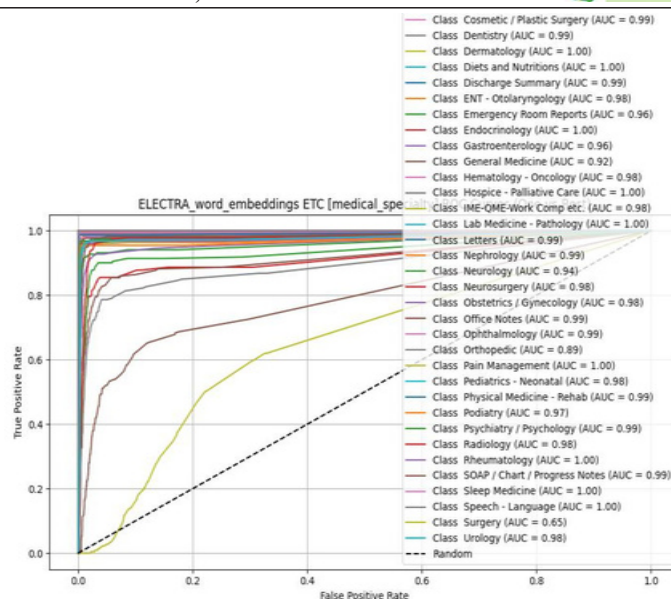


Figure 4: ROC Curve obtained using ELECTRA-WE for Proposed ET.

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.

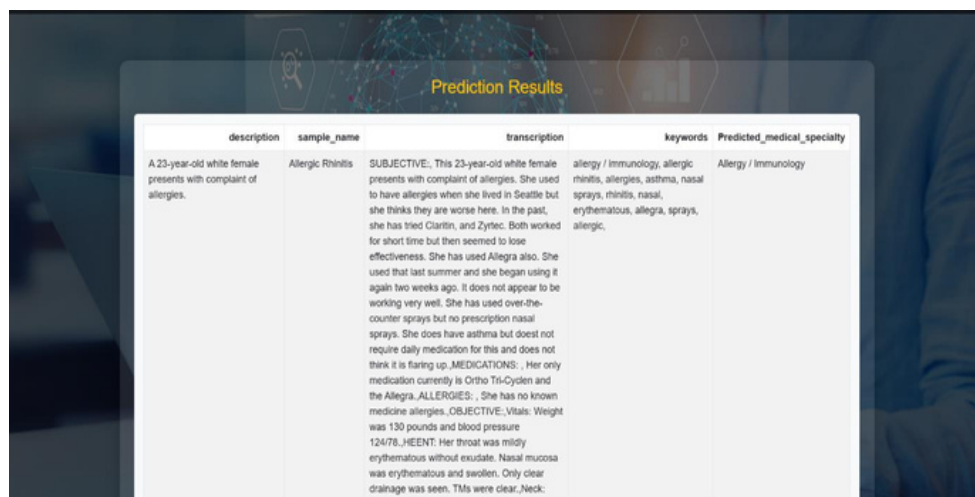


Figure 5: Real time predictions of Clinical NLP Trails.

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
----	------	------	------	------

The table 1 presents a comparative evaluation of four algorithms, namely AB, TT, RF, and ET, based on Accuracy, Precision, Recall, and F1-Score. Among all methods, ET demonstrates the best performance with all four metrics reaching 99.0%, indicating highly consistent and near-perfect classification capability. RF follows as the second-best model, achieving 86.3% Accuracy, 84.4% Precision, 86.3% Recall, and an F1-Score of 85.2%, which reflects strong and reliable performance across all evaluation measures. In contrast, TT shows moderate performance with 20.1% Accuracy, 15.6% Precision, 20.1% Recall, and 15.3% F1-Score, suggesting limited effectiveness in handling the classification task. AB records the lowest performance among all models, with 13.9% Accuracy, 12.8% Precision, 13.9% Recall, and 11.0% F1-Score, indicating poor predictive capability. Overall, the results clearly highlight the superiority of ET over the other algorithms, demonstrating its effectiveness and robustness in achieving high classification accuracy and balanced performance across all evaluation metrics.

5. CONCLUSION

This study introduces a robust framework for clinical text classification, leveraging ELECTRA-based word embeddings and ensemble learning to accurately predict medical specialties from unstructured MT Samples reports. A comprehensive preprocessing pipeline encompassing text normalization, tokenization, stop word removal, and lemmatization effectively eliminated noise and standardized input, while SMOTE addressed severe class imbalance to ensure equitable representation of rare specialties. Among the evaluated models, the proposed ET achieved near-perfect performance with 99.0% accuracy, precision, recall, and F1-score, as confirmed by confusion matrices and multi-class ROC curves ($AUC \geq 0.98$), significantly outperforming AB (13.9%), TT (20.1%), and RF (86.3%). ET exceptional results stem from extreme randomization and deep decision paths that fully exploit rich contextual embeddings. The framework offers high generalizability, reproducibility, and clinical utility, paving the way for automated disease-specialty mapping in real-world transcription 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 Khudhairi, 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] Bollegala, D., Atanasov, V., Maehara, T., & Kawarabayashi, K. (2018). Classinet—predicting missing features for short-text classification. *arXiv:1804.05260*.

- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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
- [12] 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>.
- [13] 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)
- [14] 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.
- [15] Fertig N, Domsic RT, Rodriguez-Reyna T, Kuwana M, Lucas M, Medsger TA Jr, et al. Anti-U11/U12 RNP antibodies in systemic sclerosis: a new serologic marker associated with pulmonary fibrosis. *Arthritis Rheum*