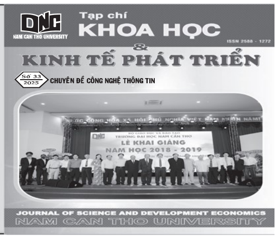
Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn


Tạp chí
KHOA HỌC
&
KINH TẾ PHÁT TRIỂN
CHUYÊN ĐỀ CÔNG NGHỆ THÔNG TIN
JOURNAL OF SCIENCE AND DEVELOPMENT ECONOMICS
NAM CAN THO UNIVERSITY

Enhancing security and scalability in electronic medical records (emr) management: Integrating blockchain and machine learning

Tran Thanh Nam[1*], Truong Hung Chen[1], Duong Van Phieu[2]

[1]*Faculty of Information Technology, Nam Can Tho University*

[2]*Faculty of Medicine, Nam Can Tho University*

[*]*Corresponding author: Tran Thanh Nam (email: ttnam@nctu.edu.vn)*

Received: 30/12/2024

Revised: 25/2/2025

Accepted: 20/2/2025

***Keywords:** blockchain, electronic medical records, healthcare, knowledge system, machine learning*

***Từ khóa:** blockchain, hệ thống tri thức, hồ sơ y tế điện tử, học máy, y tế*

ABSTRACT

The management of electronic medical records (EMRs) is a crucial yet challenging aspect of modern healthcare. In today's world, data security, privacy, and transparency are of utmost importance. This research proposes a blockchain-based knowledge system for EMR management, utilizing the Hyperledger Fabric framework and machine learning (ML) algorithms. By combining the immutable and decentralized nature of blockchain with the predictive capabilities of ML, this system aims to support medical diagnostics and decision-making, with a specific focus on stroke prediction. Experimental results using over 130,000 real-world anonymized EMRs have demonstrated that this system significantly improves data sharing among healthcare facilities while maintaining security and data integrity. The system's decision tree model achieved an impressive 85% accuracy in predicting stroke risk, highlighting the potential for blockchain-ML integration to revolutionize healthcare management and diagnostics.

TÓM TẮT

Quản lý hồ sơ y tế điện tử (EMR) là một khía cạnh quan trọng nhưng đầy thách thức trong ngành y tế hiện đại, nơi mà bảo mật dữ liệu, quyền riêng tư và tính minh bạch được đặt lên hàng đầu. Nghiên cứu này đề xuất một hệ thống tri thức dựa trên blockchain để quản lý EMR sử dụng khung Hyperledger Fabric kết hợp với các thuật toán học máy (ML). Bằng cách tận dụng tính chất bất biến, phi tập trung của blockchain và khả năng dự đoán của ML, hệ thống này hỗ trợ chẩn đoán y khoa và ra quyết định, với trọng tâm cụ thể là dự đoán đột quỵ. Kết quả thực nghiệm với hơn 130.000 EMR ẩn danh *từ thực tế cho thấy hệ thống này cải thiện đáng kể việc chia sẻ dữ liệu giữa các cơ sở y tế đồng thời duy trì tính bảo mật và toàn vẹn dữ liệu. Mô hình cây quyết định của hệ thống đạt độ chính xác 85% trong việc dự đoán nguy cơ đột quỵ, nhấn mạnh tiềm năng của sự kết hợp blockchain và ML trong việc cách mạng hóa quản lý và chẩn đoán y tế.*

1. INTRODUCTION

Healthcare systems around the globe are increasingly reliant on digital technologies to manage patient data. Electronic medical records (EMRs) are pivotal to this transformation, enabling efficient data sharing between healthcare providers, enhancing diagnostic accuracy, and improving patient outcomes. However, these digital systems introduce significant concerns related to privacy, data integrity, and scalability. Traditional centralized systems, whether cloud-based or on-premise, are vulnerable to security breaches, unauthorized access, and data manipulation, which undermine patient privacy and trust in healthcare systems [6].

Blockchain technology offers a decentralized and immutable framework to address these challenges. Initially popularized by cryptocurrencies such as Bitcoin, blockchain has found applications across various industries, including supply chain management, finance, and healthcare [22]. In healthcare, blockchain's capacity for secure, transparent, and auditable data storage can be harnessed to protect sensitive medical information. Moreover, its decentralized nature ensures that no single entity has overarching control, reducing the risk of data tampering or misuse [10].

In parallel, machine learning has revolutionized data analysis in healthcare. By training algorithms on large datasets, ML models can predict disease outcomes, optimize treatment plans, and identify risk factors with greater precision than traditional methods [20]. Combining blockchain with ML, especially in the context of EMR management, offers a transformative approach to secure data storage while simultaneously enabling predictive healthcare applications.

This paper presents a blockchain-based system for EMR management that incorporates ML to predict stroke risk. The system is built on Hyperledger Fabric, a permissioned blockchain framework designed for enterprise applications. A decision tree algorithm was applied to a dataset of over 130,000 anonymized EMRs to predict stroke occurrence. This research explores the architecture, performance, and implications of such a system in modern healthcare, examining its potential to enhance security, scalability, and predictive analytics.

1.1 Blockchain in healthcare

Blockchain technology has emerged as a secure solution for data management across various sectors, with healthcare being one of the most promising areas of application [9]. A key challenge in healthcare is the secure sharing of EMRs between hospitals, clinics, and other healthcare institutions. Traditional systems rely on centralized databases, which are susceptible to breaches and hacking attempts. Blockchain's distributed ledger system mitigates these risks by decentralizing data control, ensuring that no single entity has the authority to alter patient records without consensus from all participants in the network [19].

For instance, research by Azaria et al. (2016) [2] demonstrated how blockchain could be used to securely store and share patient data without relying on a central authority. Similarly, Zyskind et al. (2015) [23] introduced a decentralized framework that utilizes blockchain to protect personal data in a variety of contexts, including healthcare. Blockchain's transparency also ensures that all transactions involving patient data, such as updates or queries, are recorded and can be audited by authorized parties, which improves trust in healthcare systems [22].

However, while blockchain addresses many security and privacy concerns, its integration into healthcare has not been without challenges. The most significant issues include scalability, transaction speed, and the initial cost of blockchain deployment. These limitations must be addressed before blockchain can be widely adopted in the healthcare sector [8].

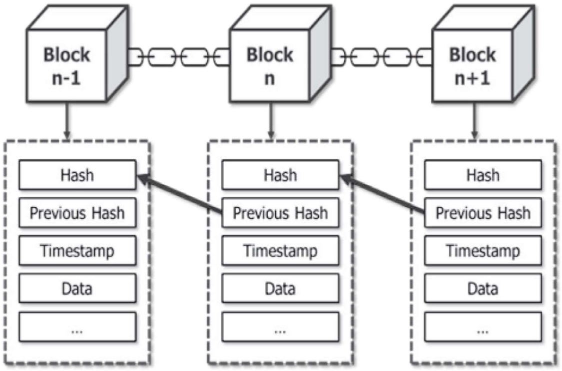


Figure 1. Structure of the blockchain

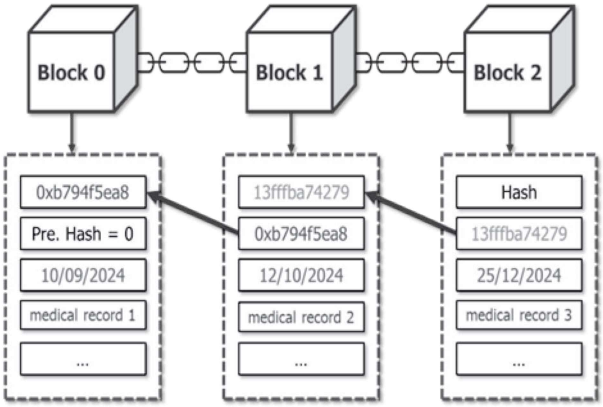


Figure 2. Example structure of the blockchain

1.2 Hyperledger Fabric: A permissioned blockchain for healthcare

While public blockchains like Bitcoin and Ethereum are fully decentralized and open to anyone, permissioned blockchains offer more controlled environments suitable for enterprise applications, including healthcare [21]. Hyperledger Fabric, developed by the Linux Foundation, is a permissioned blockchain that supports modular architectures, allowing organizations to customize their blockchain networks according to their specific needs [1].

In healthcare, Hyperledger Fabric’s permissioned structure ensures that only authenticated participants—such as hospitals, doctors, and patients—can access specific data (Azaria et al., 2016) [2]. This granular control over access rights is critical in environments that demand high levels of privacy and security. Moreover, smart contracts (or chain code in Hyperledger Fabric) allow healthcare organizations to automate access policies and ensure that transactions follow predefined rules. These features make Hyperledger Fabric a robust choice for managing EMRs in a blockchain ecosystem [1].

Studies by Yue et al. (2016) [24] highlight how Hyperledger Fabric can be used to develop secure EMR systems that maintain patient privacy while enabling authorized sharing of medical records. Additionally, projects like MedRec have demonstrated the practical application of blockchain in managing healthcare data, underscoring the potential of permissioned blockchains like Hyperledger Fabric to support secure, scalable EMR systems (Azaria et al., 2016) [2].

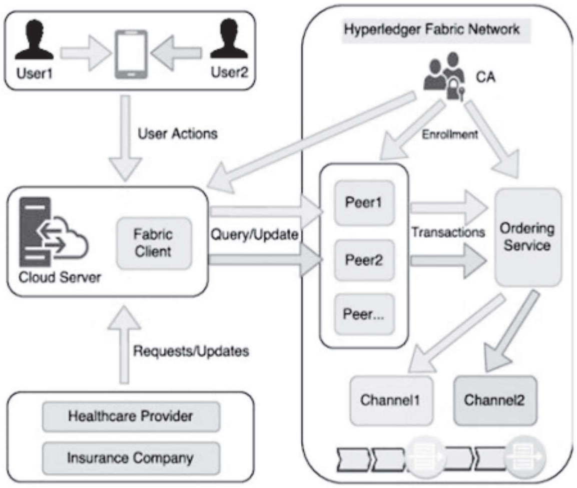


Figure 3. Blockchain Hyperledger Fabric

(Source: https://hyperledgerfabric.readthedocs.io)

1.3 Machine learning in healthcare

Machine learning has gained prominence in healthcare for its ability to analyze complex datasets and generate predictive insights. Supervised learning models, such as decision trees, random forests, and support vector machines, are frequently used for disease prediction and risk assessment [18]. ML algorithms can analyze EMRs to identify patterns and correlations that might be missed by traditional statistical methods. This is particularly valuable in predicting diseases like stroke, heart disease, and diabetes, where early detection is critical for successful treatment [20].

In recent years, there has been a growing interest in integrating ML with blockchain systems. Blockchain provides a secure, immutable environment for storing medical data, while ML algorithms use this data to make real-time predictions about patient outcomes [16]. For instance, Patel [17] discusses how ML models trained on blockchain-protected datasets can significantly improve diagnostic accuracy by ensuring that the data used is tamper-proof and of high quality.

The use of decision tree algorithms, as applied in this research, is well-suited for medical diagnosis due to their simplicity and interpretability. Decision trees are capable of handling both categorical and continuous data, making them versatile for predicting diseases based on various patient attributes [4]. However, decision trees are prone to overfitting, which necessitates careful tuning of the model, as well as validation techniques to ensure generalizability across different patient populations (Dietterich, 2000) [25].

2. RESEARCH METHODS

2.1 Blockchain framework for EMR management

The proposed system utilizes Hyperledger Fabric to securely manage and share EMRs. The system's architecture is divided into four layers to optimize data management, security, and scalability:

Data-as-a-Service (DaaS): This layer is responsible for collecting, storing, and retrieving EMRs on the blockchain. Data is stored as immutable blocks linked through cryptographic hashes, ensuring that once a record is added, it cannot be altered or deleted without consensus.

Information-as-a-Service (IaaS): Processes data to extract useful insights, including patient demographics, diagnostic results, and treatment history.

Knowledge-as-a-Service (KaaS): In this layer, machine learning algorithms are applied to processed data. For this study, a decision tree model was used to predict stroke risk based on patient symptoms and medical history.

Application Layer (APP): Provides healthcare providers with secure access to patient records and ML-based predictive insights through web and mobile applications. The application layer ensures that users can interact with the system in real-time, querying patient records and receiving stroke risk assessments instantly.

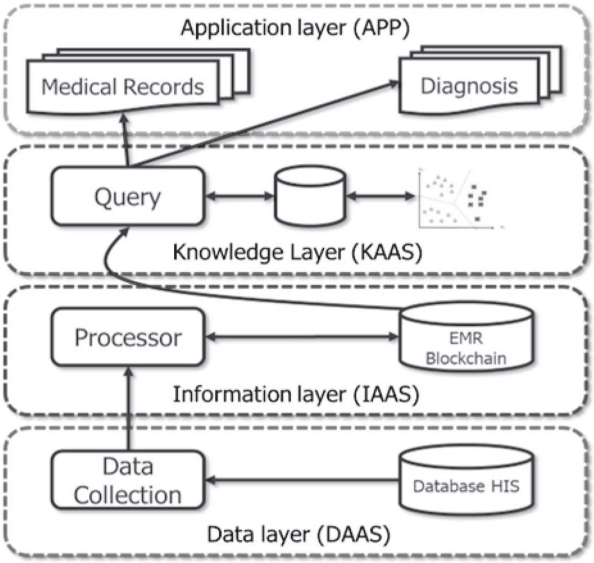


Figure 4. Overview architecture for EMR blockchain

2.2 Smart contracts for access control

Smart contracts in the Hyperledger Fabric network ensure that only authorized personnel can access or modify EMRs. These contracts manage role-based access control, ensuring that healthcare providers can only view records relevant to their responsibilities. For example, a cardiologist may be granted access to heart-related medical data, while a general practitioner might have broader access to other types of records [11].

Smart contracts also provide an audit trail, which records all interactions with patient data. This ensures transparency and accountability, as every action—whether querying, updating, or transferring records—is logged on the blockchain. This audit trail can be invaluable in case of disputes or if a breach is suspected.

2.3 Machine learning model development

The decision tree algorithm was chosen for stroke prediction due to its interpretability and efficiency in handling large datasets. The decision tree was trained on a dataset of 130,070 anonymized medical records, which included patient demographics (age, gender), medical history, diagnostic results, and specific attributes related to stroke risk, such as blood pressure, heart rate, and symptoms like dizziness or headaches.

The dataset was split into 80% for training and 20% for testing, ensuring that the model was exposed to diverse patient records during the training phase. Preprocessing steps included anonymizing the data to protect patient privacy and normalizing health-related variables to improve model performance. Apache Spark's MLlib was used to distribute the data processing and model training across multiple nodes, significantly reducing training time and improving scalability.

The decision tree model outputs a binary classification: whether the patient is at risk of stroke (label 1) or not (label 0). Features such as age, systolic and diastolic blood pressure, heart rate, and diagnosis codes (ICD) played crucial roles in the model's predictions [7]. The accuracy, precision, recall, and F1 score were calculated to evaluate the model's performance.

2.4 Experimental setup

2.4.1 Data preprocessing

The medical dataset was preprocessed using Apache Spark. Preprocessing steps included the removal of personally identifiable information (PII) to ensure patient privacy. Missing values were handled using the dropna() function in Spark, and relevant health metrics were normalized to standardize the inputs for the decision tree algorithm. For example, systolic and diastolic blood pressure values were normalized to a common scale to avoid skewing the results [12]. Additionally, the data was divided into two subsets: patient information (e.g., demographics) and medical visit data (e.g., diagnostic results). These subsets were stored in CSV format and served as input for the decision tree model.

2.4.2 Decision tree training

The decision tree model was trained on 80% of the dataset, while 20% was held out for testing. Key features included in the model were patient age, gender, blood pressure, heart rate, symptoms (as recorded by physicians or reported by patients), and final diagnostic results. The decision tree splits were based on the Gini impurity criterion, which minimizes classification error at each node [3]. To avoid overfitting, the maximum depth of the tree was restricted, and cross-validation techniques were employed. These steps ensured that the model could generalize to new data without becoming overly complex.

2.5 Blockchain configuration

Hyperledger Fabric was deployed on a network consisting of four peer nodes, each with 8-core CPUs and 32GB RAM. This configuration mimics a real-world hospital network, where multiple institutions require access to the same patient data without compromising security or data integrity. Each medical record was stored as an individual block on the blockchain, with cryptographic hashes linking the blocks together. The network used a combination of endorsement policies to ensure that only authorized nodes could validate transactions. The consensus mechanism employed was a practical Byzantine fault tolerance (PBFT) algorithm, which provided robustness against potential node failures [5].

3. RESULTS AND DISCUSSION

3.1 Blockchain system performance

The blockchain-based EMR system significantly improved data security, transparency, and access control compared to traditional centralized systems. The use of smart contracts enabled granular control over data access, ensuring that only authorized users could view or modify patient records. Furthermore, the immutability of the blockchain guaranteed that once a medical record was added, it could not be altered, reducing the risk of tampering or unauthorized modification [15]. In terms of performance, the average transaction time for querying or adding a medical record was less than 2 seconds. This low latency makes the system suitable for real-time healthcare applications, where quick access to patient data is critical [6]. The decentralized nature of the system also improved fault tolerance, ensuring that even if one node in the network failed, the remaining nodes could continue to operate without disruption.

3.2 Machine learning model performance

The decision tree model achieved an accuracy of 85% in predicting stroke risk, with a precision of 0.84 and a recall of 0.81. These results indicate that the model was effective in identifying patients at risk of stroke based on the available EMR data. The decision tree's interpretability allowed healthcare providers to understand the factors that contributed to stroke risk, which is essential for clinical decision-making [20]. The training process using Apache Spark's MLlib took approximately 15 minutes for the entire dataset, demonstrating the system's scalability and efficiency in handling large volumes of medical data. The model's performance was also evaluated using the F1 score, which reached 0.82, indicating a good balance between precision and recall [4].

3.3 Scalability and latency

The system demonstrated excellent scalability as more nodes were added to the blockchain network. The transaction time remained under 2 seconds, even when the system was processing multiple queries simultaneously. This scalability is crucial for healthcare systems that require access to vast amounts of data from multiple institutions [8]. Latency was consistently low across the network, making the system suitable for real-time healthcare applications. As healthcare providers require immediate access to patient records, the ability to query and retrieve data in less than 2 seconds is a significant advantage over traditional systems, which often suffer from delays and bottlenecks [19].

3.4 Challenges and limitations

3.4.1 Machine learning limitations

The decision tree model, while effective for stroke prediction, may not generalize well to other types of medical conditions. Decision trees are prone to overfitting, especially when dealing with complex datasets. In this study, cross-validation techniques were used to mitigate overfitting, but more advanced machine learning algorithms, such as random forests or deep learning models, may provide better results for other medical conditions (Dietterich, 2000) [25]. Additionally, the performance of the machine learning model is highly dependent on the quality and quantity of the input data. In healthcare, obtaining high-quality labeled data can be challenging, particularly when dealing with sensitive patient information. Future work could focus on improving data collection processes to enhance the model's performance and ensure that it generalizes well to different patient populations [18].

3.4.2 Blockchain limitations

While the blockchain-based EMR system offers substantial benefits in terms of security and transparency, it also presents several limitations. The immutability of blockchain records, while valuable for ensuring data integrity, can pose challenges in cases where errors are made in medical records. If a healthcare provider makes a mistake in entering patient data, it cannot be corrected directly. Instead, new records must be added to amend the error, which can lead to data redundancy and complexity [13]. Furthermore, the deployment and maintenance of a blockchain network require significant resources. Smaller healthcare institutions may struggle to implement blockchain solutions due to the cost and technical expertise required. This could hinder the widespread adoption of blockchain in healthcare, particularly in developing regions with limited IT infrastructure [22].

3.5 Future directions

There are several avenues for future research based on the findings of this study. First, the system could be extended to include more advanced machine learning algorithms, such as random forests, gradient boosting, or deep learning models, to improve prediction accuracy for a broader range of diseases. Deep learning models, in particular, could be beneficial in analyzing more complex datasets, such as medical images or genomic data [14].

Second, the blockchain network could be expanded to include more healthcare providers, enabling greater data sharing and collaboration across institutions. This would enhance the utility of the system in supporting medical diagnostics and decision-making across a wider range of healthcare facilities [21].

Finally, future research could explore the integration of additional data sources, such as wearable devices or real-time monitoring systems, to provide more comprehensive patient health profiles. By incorporating real-time data into the blockchain-ML system, healthcare providers could make more informed decisions, leading to better patient outcomes [20].

4. CONCLUSION

The integration of blockchain technology and machine learning in healthcare offers a transformative approach to managing EMRs securely and efficiently. This research demonstrates the potential of a blockchain-ML hybrid system to enhance healthcare outcomes, particularly in predicting stroke risk. By leveraging the security and transparency of blockchain and the predictive capabilities of machine learning, the proposed system addresses key challenges in EMR management, such as data privacy, scalability, and predictive accuracy. The experimental results highlight the system's ability to provide secure, scalable, and accurate predictions for stroke risk, making it a valuable tool for healthcare providers. While there are challenges associated with deploying blockchain and machine learning in healthcare, the benefits in terms of data security, privacy, and clinical decision support outweigh these limitations. Future work will focus on expanding the system's capabilities by incorporating more advanced machine learning models, additional healthcare providers, and real-time data sources. By continuing to develop and refine blockchain-ML integration, this research contributes to the ongoing effort to revolutionize healthcare data management and improve patient outcomes.

REFERENCES

[1] Androulaki, E., et al. (2018). Hyperledger Fabric: A Distributed Operating System for Permissioned Blockchains. *Proceedings of the Thirteenth EuroSys Conference.*

[2] Azaria, A., et al. (2016). MedRec: Using Blockchain for Medical Data Access and Permission Management. *2016 2nd International Conference on Open and Big Data (OBD).*

[3] Breiman, L., et al. (1984). *Classification and Regression Trees.* Routledge.

[4] Caruana, R., & Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. *Proceedings of the 23rd International Conference on Machine Learning.*

[5] Castro, M., & Liskov, B. (1999). Practical Byzantine Fault Tolerance. *OSDI '99:*

Proceedings of the Third Symposium on Operating Systems Design and Implementation.

[6] Dimitrov, D. V. (2019). Blockchain Applications for Healthcare Data Management. *Healthcare.*

[7] Duda, R. O., et al. (2001). *Pattern Classification*. Wiley.

[8] Esposito, C., et al. (2018). *Blockchain: A Panacea for Healthcare Cloud-Based Data Security and Privacy.* IEEE Cloud Computing.

[9] Gordon, W. J., & Catalini, C. (2018). Blockchain Technology for Healthcare: Facilitating the Transition to Patient-Driven Interoperability. *Computational and Structural Biotechnology Journal.*

[10] Gupta, M. (2020). *Blockchain for Healthcare Systems.* Springer.

[11] Gupta, S., et al. (2020). Role of Blockchain in the Internet of Medical Things: Transforming Healthcare Security. *IEEE Access.*

[12] Hastie, T., et al. (2009). *The Elements of Statistical Learning.* Springer.

[13] Khan, S., et al. (2020). Blockchain and Healthcare: A Literature Review. *IEEE Access.*

[14] LeCun, Y., et al. (2015). Deep Learning. *Nature.*

[15] Nakamoto, S. (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System.* Bitcoin.org.

[16] Nguyen, D. C., et al. (2019). Blockchain and Machine Learning for Collaborative Intrusion Detection in Healthcare Internet of Things. *IEEE Internet of Things Journal.*

[17] Patel, V. (2018). A Framework for Secure and Decentralized Sharing of Medical Imaging Data via Blockchain Consensus. *Health Informatics Journal.*

[18] Rajkomar, A., et al. (2018). Machine Learning in Medicine. *New England Journal of Medicine.*

[19] Roehrs, A., et al. (2017). Personal Health Records: A Systematic Literature Review. *Journal of Medical Internet Research.*

[20] Topol, E. J. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. *Basic Books.*

[21] Xia, Q., et al. (2017). MedShare: Trust-Less Medical Data Sharing Among Cloud Service Providers via Blockchain. *IEEE Access.*

[22] Zhang, P., et al. (2018). Blockchain Technology for Healthcare: Applications and Challenges. *Healthcare.*

[23] Zyskind, G., et al. (2015). Decentralizing Privacy: Using Blockchain to Protect Personal Data. *Proceedings of the 2015 IEEE Security and Privacy Workshops (SPW).*

[24] Yue, X., Zhang, Y., Xu, X., & Chen, G. (2016). A blockchain-based EMR sharing scheme with secure and efficient privacy-preserving for healthcare. *In Proceedings of the IEEE International Conference on Communications (ICC),* 1-6. https://doi.org/10.1109/ICC.2016.7510909

[25] Dietterich, T. G. (2000). Machine learning for biomedical applications. *In Proceedings of the 17th International Conference on Machine Learning (ICML), 302-309.* https://doi.org/10.1145/645530.655049