The Application of Machine Learning in the Medical Industry

Shouhe Chen

University of London, Data science and Business Analysis, Computer and Information Science, Singapore, 599491, China

1132848581@qq.com

Abstract. In the rapidly evolving landscape of the information age, the integration of machine learning has become indispensable within the medical industry. This essay delves into the application of machine learning within two key branches: pharmaceutical and healthcare. It explores how machine learning drives advancements in various stages of drug development, such as target identification, lead generation and optimization, and streamlining clinical trials, thereby enhancing efficiency and cost-effectiveness. Within the healthcare sector, machine learning revolutionizes traditional workflows and diagnostic methods, offering valuable guidance for medical professionals in their diagnoses. This transformative technology extends its benefits to a broad spectrum of stakeholders, including researchers, physicians, and patients, thereby significantly improving healthcare outcomes. Drawing upon a synthesis of literature on machine learning in the medical domain and insights from reports by leading healthcare companies, this essay underscores the tangible impact of machine learning. A compelling real-world case study, such as that of Flatiron Health, further illustrates the profound enhancements facilitated by machine learning in healthcare delivery. However, challenges accompany the widespread adoption of machine learning, such as determining its appropriate use, addressing infrastructure limitations, ensuring the quality of training data, and mitigating issues of overfitting and underfitting. Despite these challenges, research indicates that the overarching benefits of machine learning outweigh the drawbacks. As exploration, adaptation, and cross-industry learning continue to shape the evolution of machine learning in medicine, its potential to revolutionize the field remains promising.

Keywords: Artificial Intelligence, Machine learning, Medical, Pharmaceutical industry, Healthcare industry, Innovation.

1 Introduction

1.1 Research Background

In today's rapidly developing information era, there is a signal that technology can change human lives to some extent, the concept of artificial intelligence which is

Z. Zhan et al. (eds.), *Proceedings of the 2024 10th International Conference on Humanities and Social Science Research (ICHSSR 2024)*, Advances in Social Science, Education and Humanities Research 858,
https://doi.org/10.2991/978-2-38476-277-4_103

known as AI has gradually drawn more and more attention from public, and this new concept is now being used into various industries by managers to different degrees. As a base of AI development, machine learning plays a significant role in the development of AI. Machine learning is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can learn from data and generalize unseen data, and thus perform tasks without explicit instructions. [1] In other words, machine learning is about how human let machine process the specific tasks automatically. Nowadays, machine learning approaches has be applied into various field including large language models, email filtering, and speech recognition. In this essay, two main domains of medical industry will be introduced and analyzed, which are pharmaceutical and healthcare industry.

1.2 Literature Review

Recently, scientists and researchers are digging into how to apply AI and machine learning into medical industry smoothly, and whether the new methods is reasonable to apply at specific stages. Recent research [2] has shown that machine learning can assist in facilitating the clinical and translational researches during covid-19 period. With the help of machine learning, the peaks and size of the pandemic globally can be predicted, and the effectiveness of some policies like social distancing to reduce the transmission of pandemic. Otherwise, machine learning changed the way of examination to test the disease, such as the modification in X-ray. This literature provides insight into the potential of machine learning to enhance various aspects of the medical industry, including new drug development, pandemic prediction, and medical examinations during special periods. As such, this essay aims to investigate the rationale behind the application of machine learning in the medical industry during normal periods. It seeks to determine whether machine learning continues to offer significant benefits when applied in the medical field under standard conditions.

1.3 Literature Sources and Research Strategy

This study focus on the reports and publications about the background and prospects of machine learning, and the how can it be applied into medical industry. Wikipedia, Google scholar, National Library of Medicine were searched, which were restrict to search the reports and article published in English. The search syntax includes the following terms: “Machine Learning”, “Artificial Intelligence”, “Medical Industry”, “Innovation”, “Deep Learning”, “New drug development”, “Healthcare”.

2 The Changes Brought to Medical Industry

The appearance of AI and machine learning not only has changed the business models, but also have changed how people work, treat their customers, and improve the innovation. Today, AI and machine learning have been introduced into the medical industry, and have some beneficial impacts on various branches of medical industry, the

beneficial field includes pharmaceutical industry, clinical medicine, and the communication management.

2.1 The Changes in Pharmaceutical Industry

The changes brought by pharmaceutical industry can be considered the most dramatic. Through the traditional drug development process, the researchers have to spend a lot of time and money on the thousands to millions of compounds optimization, but only a few results that may relevant to the targets are yielded. According to the research, only about 3-5% trial enrollment can get the desired results which is really beneficial for the biological research and analysis. [3] However, the average cost of developing a new drug was estimated to be $2.6 billion in 2013, while the success rate was only about 12%. [4] Which means the traditional drug development is time consuming and costly but inefficient. The application of AI and its subset, machine learning, change the drug development process including the target identification, lead generation, lead optimization, preclinical trials, and clinical trails.

2.1.1 Target Identification

The term 'target identification' is mainly about identifying whether a disease is 'druggable'. Machine learning creates a brand new path for identifying and validating the new drug targets. The AI algorithms can pinpoint potentially therapeutic targets with higher precision by sifting through voluminous datasets that span genomics, proteomics, metabolomics, and transcriptomics. [5]

Nowadays, researchers have already set about establishing machine learning model using deep learning to help them model the molecular consequences of genetic variations, this kind of model is able to access the data include human gene sequence, protein structure, molecular structure to make some prediction about the interactions between drugs and their targets. Also, machine learning algorithms can analyse vast amounts of biological, chemical, and pharmacological data to identify patterns and relationships that may not be immediately apparent to human researchers, but provide an insights for the further research. In recent years, deep learning model such as AlphaFold, RoseTTAFold, CNNs, and RNNs are applied to predict protein structure and design approach.[6] Predicting protein structures helps in understanding their functions and identifying potential binding sites for drug molecules. Although some deep learning can now only be applied on pure protein system, some steps have already been made to explore the machine learning in target identification. Otherwise, before the wide population of AI, it is always hard to identify the patients-specific drug targets for individual patients or subpopulations. But there is machine learning algorithms that can figure this problem based on genetic, genomic, and clinical data in a cost saving manner. The model can predict which drugs are most likely to be effective for individual patients or subpopulations, which can lead to more personalized and effective therapies.

2.1.2 Lead Generation and Optimization

Hit to lead (H2L) also known as lead generation is a stage in early drug discovery where small molecule hits from a high throughput screen (HTS) are evaluated and undergo limited optimization to identify promising lead compounds.[7] Drug lead optimization is the process of refining and improving the properties of a lead compound identified during the drug discovery phase to develop a candidate drug with enhanced efficacy, safety, and pharmacokinetic properties. In other words, lead generation is trying to generate a drug compounded by different kind of molecules, and lead optimization is to make some modification to the new drug.

Machine learning algorithms can help to do virtual screening, which can perform virtual screening of large compound libraries to identify molecules with the potential to bind to a target of interest. These algorithms use various molecular descriptors and structural features to predict the likelihood of a compound interacting with the target, thus narrowing down the pool of candidates for further experimental validation. Also, it can predict the absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of compounds based on their chemical structures. By integrating these predictions into the lead generation process, researchers can prioritize compounds with favorable ADMET profiles and minimize the likelihood of toxicity or other adverse effects in later stages of drug development.

2.1.3 Clinical Trails

With the help of machine learning algorithms, researchers can filter the patients who fulfill the test standard of the new drug. The algorithm can analyse patients' features based on their gender, age, genetic gene and medical history, then combine the information of the new drug to predict whether the patient is able to accept the test that have the lowest probability of failure. By leveraging data from previous clinical trials and real-world evidence, these models can provide insights into patient outcomes and help optimize treatment strategies, dose selection, and patient stratification in clinical trials.

It is critical that the new drug trails will not hurt the patients. There are professionals who will record the real-time status and upload them into the database. Then machine learning algorithms can monitor safety signals and adverse events in real-time by analyzing the uploaded data combined with patients' electronic health records, clinical trails in the database or social media platform. By identifying potential safety issues early, these models can enable proactive safety monitoring and risk mitigation strategies, therefore improving patient safety and regulatory compliance.

2.2 The Changes in Healthcare Industry

In healthcare industry, each doctor, physician, or even nurse have to face a bunch of patient's information everyday and these information are usually unique and timely as there are new patients who visit hospitals and clinics everyday. But the thing is that each patients' information such as their examination results, diagnosis plans, recommendations for Rehabilitation are equally valuable for these medical practitioner and are required to be treat carefully. These tasks seem that will not be handled using

human hand now as the medical data is exploded enormously. Some ERP system have been developed and applied in hospitals and clinics to help figure this problem out. However, the application of the ERP system is not enough, far from enough to relieve the pressure brought by data explosion on staff, so machine learning algorithms are introduced while machine learning can change the workflow of those doctors and physicians to make them work more efficiently, so that they can have more time to stay with the patients and provide them a better care, which is also the prior core value of healthcare.

2.2.1 Application of Machine Learning in Healthcare

Flatiron Health, a healthcare company that experts on improving cancer treatment and advanced research, [8] use machine learning algorithms in their data curation and diagnosis stage. Flatiron can touches the data of more than 2 million patients annually as they have the permission to obtain the data from their cooperated hospitals and clinics, [9] and these data are always unstructured, some data can be only one signal from the examination image or research experiments. It is hard to identify them correctly using human eyes. The researchers use the algorithms and establish machine learning models then train them that the model can identify various kind of information produced by different research and transform them into the data that can be read and processed by other system automatically. This process can also finish some tasks like identifying the extreme values that may disturb the further experiment, and clear some 'noise' that can influence the accuracy of the research results, which the whole process included the transforming is considered as 'data curation'.

In addition, an innovation system called Flatiron Assist is created to optimize the diagnosis process. After those data about patients are transformed well and already made sense, then the data can be used to analyze the situation for each individual patients. Flatiron Assist have an interface that enable manipulators enter the brief explanation of the specific symptom after diagnosis, usually a few lines of text, it will began to search the similar symptom cure plan based on the diagnosis history stored in the database, then it will give some recommended cure methods or drugs to the doctors, and doctors can make some modification to the recommendation and form the final diagnosis plan. The final plan will be fed back to the system as the new training data to improve Flatiron Assist. During the process, machine learning plays a role as the foundation of the whole system, the system will not operate normally.[10] The feature that Flatiron Assist can reuse the produced data to improve the algorithm can be considered as a reflection of machine learning's characteristic.

The thing needed to pay attention is that although some application of machine learning that helps optimize the diagnosis process has been successful in Flatiron, it is still immature compared with the traditional diagnosis, the experience is helpful in medical field after all. The success in Flatiron can be treated as a good example for the managers in medical industry to learn from.

3 The Challenges of Machine Learning in Medical Industry

3.1 Whether to Use Machine Learning

Before doing any task, the preparation step is very critical. The goals of the task, the sources will be used, the time frame, and the detail steps to do in every stage of the task are supposed to be set up during the preparation. The thing is that it is not necessary to apply machine learning for every task due to the cost. Managers need to evaluate which part should be solved with the help from machine learning. Automation is unnecessary when the task is done frequently and it is easy to predict the outcome because the variable is almost unchanged. However, some complicate task need further inspection before the automation. While Machine learning can definitely help automate some processes, not all automation problems require Machine learning.[11] Machine learning can be a good support to help optimize the results and give some prediction. The users input the data that have different features as more as possible to predict the selected feature more precisely. In medical industry, doctors can use EHR(Electronic Health Record) which is a enterprise system that collects patients' information to help them make better decisions. The system can be considered as an application of machine learning. Some algorithms are set in the system once which the key word or short form of the symptom and some recommended therapeutic methods or drugs will be provided based on the same symptom diagnosis history. However, machine learning can only provide a predicted result which means the result is not determined to be 100% accurate. Although the medical infrastructure looks more advanced, for example, patients can collect their X-ray results with some explanations, which is an application of machine learning, and physicians or doctors will diagnose referring to this explanation. Compared to the judge by doctors themselves, this kind of explanation can be a disturbance while doctors making decisions as they may not be accurate. A small alternation of the information may influence the cure. Hence, managers need to consider whether they can accept the inaccuracy of the results produced by the automation and prediction and abandon the use of machine learning, or they need to figure out how to make the automation and prediction processes more precise.

3.2 Inadequate Infrastructure

Machine learning require a vast amount of data churning capabilities, which provide some requirement to the organizations that they must have a adequate infrastructure to run the machine learning model. The infrastructure includes network, data storage can influence the AI development.[12] Unlike data processing, managers can get the aids from cloud computing platform, even if an artificial intelligence is developed successfully, the presence of AI-ready infrastructures that can run the AI successfully have to be fulfilled while the legacy system and infrastructure that can not support the workload may force the AI to buckle to a degree.

The problem that medical industry facing is also inadequate infrastructure. Because the concept of AI, machine learning, automation is just brought to public in last decade,

and medical infrastructure is usually built with the development of the country, some European and Asian countries which can afford to develop AI are almost finished their medical infrastructure construction. It is hard to totally change the modern infrastructures to the next-level infrastructures, the things like the money and time cost, how to establish the new infrastructures while do not disturb normal operation of the whole medical system, and how to promote the new AI to medical industry participants smoothly, are required to be considered.

3.3 The Quality of Data

The most important thing for machine learning is the data no matter what the purpose or the kind of machine learning model is. Machine learning is a process that people train the machine to do the task automatically that they want, and the training resource can only be different type of data, including text and numeric. Thus, the quality of data can straightly affect the performance of the machine learning processes. However, except for the distracting information that would have been there anyway, some information can be created easily with the help of generative AI,(Generative AI models learn the patterns and structure of their input training data and then generate new data that has similar characteristics [13]) which means the input data can be more inaccurate than before.

Otherwise, medical data are always be unstructured, the process of transforming them into the data in form that can be read and 'acknowledged' by the machine learning model is also significantly important. In medical industry, the information about patients including their medical history, basic personal information, and some body examination results are usually managed to be used as the source of machine learning training data to process different tasks. But it is hard to obtain the information from image like the nuclear magnetic resonance image. Before set them as the training data, some preprocessing is necessary for these image, experts are supposed to classify the image and analyse the information contained in the image somehow, this process always takes quiet a long time. However, as the improvement of the AI field, some information in the image can now be read by machine directly. Managers within medical industry still need to put some effort to prepare the data.

3.4 Avoid Overfitting and Underfitting

Overfitting is an undesirable machine learning behavior that occurs when the machine learning model gives accurate predictions for training data but not for new data. [14] The reason of overfitting can be that the training data set is too small or not sufficient. Lack of training data may cause the insufficient training of machine learning model, which means the model can not identify most features of the targets. For example, the training data only provide the people that have male characteristics so that the model can not identify most of female. Otherwise, the 'noise' in training data, too much similar data can also be the reasons of overfitting. Too much similar data will cause that the model performs super good when training but can not provide effective results of new data.

Underfitting occurs when a model is too simple and is unable to properly capture the patterns and relationships in the data. This means the model will perform poorly on both the training and the test data.[15] The reason of underfitting is always that there are not enough parameters or the machine learning model is not strong enough for the given data set. Managers need to determine the proper numbers of feature of the data. The methods of resolving underfitting can be opposites of overfitting, which is increasing the model complexity, for instance, the increase of the model complexity for k-NN model can be the increase of the numbers of neighbors which is known as k.

Overfitting and underfitting will both lower the performance of machine learning model, and it is time consuming once these problems occur. Hence, managers are advised to make sufficient preparation of the training data set and modify the model as the model is developed simultaneously.

4 Limitations

This essay has several limitations that need to be acknowledged. Firstly, there may be inadequate analysis due to the possibility of missing relevant literature during the research process. Despite efforts to comprehensively gather information, oversights or omissions are inevitable, potentially limiting the depth of analysis.

Secondly, defining "normal" conditions globally poses a challenge, as various factors such as economic fluctuations or geopolitical conflicts can influence data patterns. It's difficult to isolate the impact of machine learning within a strictly defined normalcy, as external variables may skew results and interpretations.

Lastly, the availability of information within the healthcare industry is constrained by proprietary considerations, with certain data withheld as business secrets by companies. This limitation obstructs a complete understanding of industry dynamics and may lead to incomplete or skewed analyses, particularly when assessing the effects of machine learning in healthcare.

Acknowledging these limitations is crucial for maintaining a balanced perspective and understanding the nuances inherent in evaluating the role of machine learning in the medical industry. Efforts to mitigate these limitations, such as comprehensive literature reviews, contextual understanding of global conditions, and transparency in data sharing within the healthcare sector, are essential for fostering more robust and accurate analyses in the future.

5 Conclusion

In conclusion, the integration of machine learning into the medical industry under normal conditions presents both advancements and challenges. Despite the hurdles, we believe these challenges are surmountable to some extent, and certain benefits persist beyond the pandemic period, potentially expanding further with ongoing improvements in machine learning technology.

Prior to implementing machine learning for specific tasks, managers must carefully assess at which stage it is necessary. This decision should not solely hinge on

cost-saving measures but also on whether human performance can be surpassed by AI algorithms for the given task. Paramount among these considerations is the quality of data. The efficacy of machine learning models is directly impacted by the quality of training data, yet preprocessing this data presents formidable challenges in identifying the most suitable data. Moreover, striking a balance in the quantity of training data to prevent overfitting or underfitting poses additional complexities. Furthermore, inadequate infrastructure in many developing nations impedes the advancement of higher-level technologies reliant on AI and machine learning algorithms.

Nevertheless, the potential improvements brought about by machine learning remain substantial, motivating stakeholders to address the aforementioned obstacles. In the pharmaceutical industry, machine learning revolutionizes the new drug discovery process across various stages, from target identification to lead optimization and clinical trials. By predicting outcomes for different molecular combinations and optimizing the most viable options, machine learning significantly expedites and economizes the traditional drug discovery process, also mitigating trial risks by facilitating patient monitoring and filtering prior to trials. Similarly, within the healthcare sector, machine learning streamlines workflow and enhances diagnostic efficiency by supporting decision-making and automating data processing.

However, managers in the medical industry grapple with challenges related to the widespread adoption and precision enhancement of AI and machine learning technologies at each stage. Consequently, while the application of machine learning has demonstrated promising impacts, there remains a substantial journey ahead for managers and practitioners in the medical field.

References

1. Wikimedia Foundation. (2024, March 6). Machine learning. Wikipedia. https: //en. wikipedia.org/wiki/Machine_learning.
2. JMIR Med Inform (2021, Jan 11). Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review https: //www. ncbi. nlm. nih. gov/ pmc/ articles/ PMC7806275/.
3. Barriers to patient enrollment in therapeutic clinical trials. (n.d.). *https: // www .fightcancer.org/sites/default/files/National%20Documents/Clinical-Trials-Landscape-Report.pdf.*
4. RW, D. J. H. (n.d.). Innovation in the pharmaceutical industry: New estimates of R&D costs. Journal of health economics. https://pubmed.ncbi.nlm.nih.gov/26928437/.
5. I., F. (2023, August 26). How AI is transforming the Pharmaceutical Industry. LinkedIn. https://www.linkedin.com/pulse/how-ai-transforming-pharmaceutical-industry-francis-nwakire-md-mba.
6. Professor Ewan BirneyEMBL Deputy Director General and EMBL-EBI Director, & Professor John McGeehanProfessor of Structural Biology and director for Centre for Enzyme Innovation (CEI) the University of Portsmouth. (2022, July 28). Alphafold. Google DeepMind. https://deepmind.google/technologies/alphafold/.
7. Wikimedia Foundation. (2023, December 9). Hit to lead. Wikipedia. https: //en. wikipedia.org/wiki/Hit_to_lead.

8. Patyal, S. (2018, April 9). Flatiron Health – solving cancer through data analytics. Digital Innovation and Transformation. https: //d3. harvard. edu/ platform -digit /submission/flatiron-health-solving-cancer-through-data-analytics/.
9. Patyal, S. (2018, April 9). Flatiron Health – solving cancer through data analytics. Digital Innovation and Transformation. https: //d3. harvard. edu/ platform- digit/ submission/ flatiron-health-solving-cancer-through-data-analytics/.
10. Patyal, S. (2018, April 9). Flatiron Health – solving cancer through data analytics. Digital Innovation and Transformation. https: //d3. harvard. edu/ platform- digit/ submission/ flatiron -health-solving-cancer-through-data-analytics/.
11. Data Center news and Trending Topics. 2024 Data Center News, Cloud and Technology Articles. (n.d.). https: //www. datacenters. com/ news/ the- crucial- role- of- modern-I nfrastructure-in-ai-implmentation.
12. Inc., P. I. (n.d.). 5 common machine learning problems & how to solve them. ProV International - The Best Technology Consulting Firm. https: //www. provintl. com/ blog/ 5-c ommon-machine-learning-problems-how-to-beat-them
13. Wikimedia Foundation. (2024b, March 7). Generative Artificial Intelligence. Wikipedia. https://en.wikipedia.org/wiki/Generative_artificial_intelligence.
14. What is overfitting? - overfitting in machine learning explained - AWS. (n.d.). https: //aws. amazon.com/what-is/overfitting/.
15. Dev, S. (2022, December 27). Overfitting vs underfitting. Medium. https: //medium. com/ @devsachin0879/overfitting-vs-underfitting-6a41b3c6a9ad.