INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

IJARIIT

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 11, Issue 2 - V11I2-1257)
Available online at: https://www.ijariit.com

Personalized Medicine Using AI and Genomics

Khan NavidShaba
ksaba0384@gmail.com
S. K. Somaiya College, Somaiya Vidyavihar University, Mumbai, Maharashtra

ABSTRACT

A conceptual change in healthcare, personalized medicine uses an individual's genetic profile to forecast disease risk, customize drug treatments, and increase patient outcomes. With its great genomic variety, the development of such systems is hampered in the Indian setting by the scarcity of region-specific, annotated clinical datasets. Using IndiGenome and PharmGKB as main references, this work presents a framework for combining genomic and pharmacogenomic data to enable personalized medicine. A manually built dataset with standardized notations was produced to replicate patient data due to integration difficulties between accessible datasets. This enabled the application of treatment logic and important gene mutations (BRCA1, BRCA2, TP53) based on working artificial intelligence. Future deployment with actual genomic data builds on this Streamlit-based application, which is able to predict treatments and provide health recommendations.

Keywords: *Personalized Medicine, Artificial Intelligence, Genomics, Brca1, Brca2, Tp53, Random Forest, Machine Learning, Drug Prediction, Pharmacogenomics*

INTRODUCTION

Personalized medicine is revolutionizing the health care landscape, and the "one size fits all" method is giving way to more personalized management for each person tailored to their own unique genetic composition, environmental exposure, and lifestyle behavior's. This approach is even more pertinent in managing complex diseases like cancer or heart disease, which are affected by inter-individual genetic variations critical to the progression of disease and potential pharmacological treatments.
India is especially interesting, with its diverse genetic landscape of over 4,600 ethnic groups suggesting exciting possibilities as well as challenges in the field of precision medicine. Unfortunately, the majority of genomic databases available in the world are focused on Western populations—without applicability in the Indian context. That's where IndiGenome comes in, filling this gap by offering genomic data that's specific to the region, while PharmGKB enhances this with pharmacogenomic annotations and insights into drug-gene interactions.

Although the potential for integrating these datasets is substantial, there are challenges to address: harmonization of data, compatibility of the different schema, and gaps in clinical annotations. This paper introduces a framework based on artificial intelligence (AI) to forecast patients' responses to drugs, using integrated datasets, and it provides a working prototype that was built on a well-defined integrated dataset.

Key Contributions:

i. Developed a framework for AI-driven personalized medicine using Gene Mutation profiling.
ii. Highlighting the limitations of the current genomic datasets, especially for the underrepresentation of Indian Population.
iii. Comparative analysis of IndiGenome & PharmGKB datasets and explored the challenges involved in integrating data.
iv. Creation of simulated dataset of 200 virtual patients to test and validate the proposed system in the absence of harmonized real-world data.
v. Focused on key genetic markers – BRCA1, BRCA2, & TP53- to demonstrate model's clinical relevance and real-world potential.
vi. Designed a flexible and scalable system architecture that can be extended in the future with standardized clinical and genomic datasets.

BACKGROUND AND RELATED WORK

Advances in AI for Genomics:

Recently, artificial intelligence (AI) and machine learning (ML) have revolutionized genomics, providing useful and efficient tools to analyze biological data. One of the most exciting areas within genomics is using genetic profiles to predict disease susceptibility and response to drugs. Ongoing research has included a variety of AI models for interpreting genomic sequences, particularly deep learning algorithms like convolutional neural networks (CNNs) and recurrent neural networks (RNNs). AI-enabled systems have been applied with some success to whole genome sequencing for predicting disease risk for cancer (and other diseases like Alzheimer's disease and cardiovascular diseases) based on genetic mutations. Another of the most prominent advances is the application of AI to the field of pharmacogenomics, specifically ML models that prospectively estimate the effects of genetic variants on drug metabolism and drug efficacy which form the basis for personalized pharmacotherapeutic selection. In addition, AI techniques such as reinforcement learning and generative adversarial networks (GANs) are promising in drug discovery and optimization and may bring forward more personalized therapies.

Limitations of Existing Approaches:

Although progress has been made in this direction, current AI-based genomics approaches still suffer from numerous drawbacks, particularly for populations with different backgrounds. Most AI models are trained mostly on data of Western populations which results in biased predictions when it is used on populations genetically diverse, for example, in India. This absence of diversity in training data may lead to suboptimal performance, especially for low frequency genetic variants common in non-western populations. In addition, current models have limited.

ability to seamlessly incorporate reporting of multiple genomic data categories (e.g., single nucleotide polymorphisms, structural variants, gene expression profiles) into an integrated predictive framework. Lack of fully representative, population-based clinical data sets further restricts development of pharmacogenomic models that are accurate. In the Indian context, the absence of dense genomic data considering the specific genetic differences of the Indian population hinders this problem, so the prediction of drug response in Indian patients is constrained.

Why IndiGenome and PharmGKB?

These two projects, GermlineSeq and PharmGKB are a new opportunity to break free from these constraints and integrate genomic diversity with detailed pharmacogenomic annotations. The specific purpose of IndiGenome, a novel platform that has been designed to identify the genetic variation of South Asian populations, is to offer crucial insight into the genetic composition of Indians and other South Asians who are poorly represented in the current global genomic research. Despite using IndiGenome, researchers can gain insight into the genetic variations found only in this population which is crucial to developing more accurate predictive modeling. Given the lack of clinical annotations in IndiGenome, especially related to drug responses, it suggests using a long-standing pharmacogenomic database, known as PharmGKB. PharmGKB is a rich source of information about gene-drug interactions, drug response variability, and clinical guidelines, and therefore could be useful in predicting the impact of genetic variation on drug metabolism and efficacy. By merging these two datasets, a more integrative strategy for the personalized medicine can be achieved, integrating genetic polymorphism with pharmacogenomic information, and obtaining not only a valid but also population-customized model. Such integration can be used to mitigate the current vacuum of Indian clinical datasets and provide a framework for more tailored and better treatment approaches for Indian patients.

DATASETS

IndiGenome:

The IndiGenome dataset is an important data resource implication of genomic diversity of south Asian populations, in particular Indian populations; in addition to a comprehensive list of genetic variants including SNP's and structural variants most commonly found in this population. IndiGenome hopes to address the imbalance with representation of South Asians in global genetics by presenting information relating population-specific genetic features, disease susceptibility and genetic abnormalities with greater incidence in the region. This dataset is a major step forward in building a genomic map of Indian subjects and provides pathways for more accurate disease risk assessments and personalized treatment strategies for a genetically diverse population.

PharmGKB:

PharmGKB is a pharmacogenomics database providing genetic variant-drug response relationships, with significant clinical annotation on drug-gene interactions. It provides information about the role of genetic variation in drug effects, drug metabolism, and risk of adverse drug reaction. The PharmGKB is widely used within the pharmacogenomics research and clinical decision support community, providing a number of tools and recommendations for using genetics in precision medicine. It also offers a substantial amount of clinical and research data including optimal doses of individual drugs within the context of genetic variations. However, while PharmGKB has a wealth of information, its population coverage is heavily skewed towards the Western alleles and it loses representation of non-Western genetic variants, specifically for India and South Asia. By integrating the pharmacogenomic information in PharmGKB and the heterogeneity of the genetic data available with the IndiGenome, it may be possible to address this gap and develop a more robust structure for personalized medicine in India.

CHALLENGES IN INTEGRATION

Genetic Variation Discrepancies:

One of the key obstacles to using the IndiGenome and PharmGKB data sets is genetically similar origins among Indian populations and populations represented globally within some global data sets. Indian populations have unique genetic features missing from many global genomic databases.

Variants like single nucleotide polymorphisms (SNP) and structural variants may impact susceptibility to disease, metabolism of the drug/s, and the response to therapy. Even though PharmGKB contains a vast amount of pharmacogenomic data, it mainly includes catalogue of drug-gene interactions from populations, where the genetic background is the most common in western countries. Mapping Indian specific genetic variants to the relevant clinical outcomes and drug responses found in PharmGKB is a complicated process requiring unique annotation approaches and tailored methods to address the gaps. Without adjusting for these differences, it would not be valid to apply these data sets to personalised medicine for the Indian population.

Data Format Standardization:
A further prominent issue is the structural differences in the data sets. IndiGenome and PharmGKB have different data schemas for their polymorphisms and clinical annotations, respectively, which both create challenges in combining the datatypes. To illustrate, while IndiGenome primarily represents genomic data (primarily SNPs) and providing indels and gene sequences, PharmGKB actually has clinical pharmacogenomic data as annotations (drug-effect, dose guidance, and drug-gene interaction to include some examples). The inherent dissimilarities in data format and levels of detail also require preprocessing and harmonization pipelines to create a common scale with which to use the datasets. Thus, it is important to create a coherent framework that translates and standardizes the data from two modalities to some common format in order to assure accuracy in predictive models and for everyday clinical decision-making.

Population-Specific Biases:
Artificial intelligence (AI) and machine learning (ML) models developed with mostly Western supplied data can lead to an increased risk of population-level bias in the Indian genomic space. Most of available genomic and pharmacogenomic research focus on Western populations, which may lead existing AI and ML models to miss inherent genetic variation and drug response patterns among individuals in India. This can lead to wrong prediction and a lower efficacy of individual treatment regimes. In order to reduce this, AI and ML algorithms must be carefully validated and calibrated to the genetic variability and clinical characteristics of Indian populations. AI and ML models based on population clusters must monitor and retrain algorithm whenever applied to Indian data using regionally-based databases if we are serious about the validity and utility of these models in an Indian health care system.

Add-on
Due to these challenges, including annotation and format mismatches, merging IndiGenome and PharmGKB was not feasible within this project scope. To enable practical testing, a standardized manual dataset was created.

PROPOSED FRAMEWORK

The framework was initially structured to combine the IndiGenome and PharmGKB datasets into a single pipeline to form an AI-enabled personalized medicine system. However, there were challenges associated with the practical integration of the two datasets including differences in notation representations, schema variation, and absence of certain annotations. These factors made it impossible to combine the datasets directly, compromising the practicality of that implementation within the time available.

To support this, the framework was adapted to use a manually generated dataset. The manually generated dataset consisted of 200 simulated patient records, where patient records had standard fields (i.e. age, sex and presence or absence of mutations) defined for three specific genes (BRCA1, BRCA2 and TP53). This selection of genes followed a prior negative association established with the genes above with respect to cancer risk and treatment response and had been very related to breast and ovarian cancers. Thus, the possible modelling could thus be meaningful and loosely represent real genomic relevance.

Data Processing Pipeline:
Each patient in the simulated dataset was created using a consistent format modeled from clinical genomics. Each patient has discrete values for the selected genes and demographic attributes. To allow for compatibility with supervised machine learning algorithms, the data was preprocessed by normalization followed by label encoding.

AI Model Design:
A Random Forest Classifier was identified as the primary predictive model for this assignment, due to its robustness, ease of interpretation and effective handling of small and imbalanced datasets. This model was developed around learned patterns from gene mutation indicators as well as patient demographic factors to predict the most suitable treatment option from a pre-defined list of drug and lifestyle options.

Predictive Workflow:
The processed data is divided into training and testing subsets to evaluate the performance of the model. The trained model is serialized by joblib and embedded into a web-based interface, which follows a workflow consisting of input processing, prediction generation, and output representation.

Practical Implementation with Manual Dataset:
A Streamlit web application was designed to allow user interaction with the model as described in the previous chapters. This interaction allows a user to upload patient data in .txt, .csv, or .pdf formats. The application extracts the necessary fields from the file, import data into the model, and presents a treatment prediction. In addition to the recommended treatment Letrozole (Imatinib, Tamoxifen, Olaparib), the system also provides lifestyle or dietary options under the predicted treatment.

This was not a clinical data set; to this point deliberate limit this implementation used a generated dataset which was not clinically validated, and this approach has achieved a scalable framework to which can be responsive to real genome datasets in the future, with the caveat that the datasets contain some similar standardized inputs.

ETHICAL AND SOCIAL IMPLICATIONS

While the current implementation uses a manually generated dataset that does not involve real patient information, ethical and social considerations remain central to the future deployment of AI-driven personalized medicine systems. When integrated with actual clinical and genomic data, the following concerns must be addressed:

Privacy Concerns:

A key consideration is information privacy, because genomic data is highly sensitive and should be treated with extreme sensitivity. It is key to use encryption modalities, and to comply with international data protection standards (like the General Data Protection Regulation (GDPR)) to protect individuals against privacy violations and to ensure genetic material is stored and shared in a secure manner.

Bias in Predictions:

Along with privacy, there is also the risk of bias in predictions. AI models can propagate bias, especially if the training datasets are bias-prone or imbalanced - which may result in inequitable healthcare outcomes for minority populations like in India. For instance, if the model is predicting drug responses using genetic variations associated with Indian populations in India, the model may predict poorly because there is less data supporting the eigenvectors for drug responses in Indian populations in the past literature. If the goal is to manage this Inequitable risk, then careful validation, with iterative refinements to all AI models, should be explored to ensure that all healthcare AI models are fair and accurate versus discriminatory. Regularly monitoring models and evaluating different data types can also be used to help models identify and correct bias- which could alleviate health outcomes equity thereby increasing trust and use of AI in healthcare.

APPLICATIONS AND ADVANTAGES

Practical Clinical Simulation:

The current implementation demonstrates a simulation discussing how AI can be used to support personalized treatment recommendations through structured genomic and demographic data. While the dataset used is manually developed, it represents actual clinical attributes and allows for testing and validation of the system within a safe and practical environment.

Treatment Recommendation + Lifestyle Guidance:

The AI uses the genetic markers patient specific drug treatments like BRCA1, BRCA2, TP53 and for patients at lower risk, suggested lifestyle changes (diet, exercise, screening) and expanded the usefulness of the AI systems beyond pharmacological support.

Scalability with Legitimate Data:

Although constructed utilizing a simulated dataset, the system architecture is intended to work with authentic genomic databases. Future iterations can make use of clinical records with the condition the model is structured and displayed in the same way.

Cost-Effective Development:

The proof-of-concept system, using a manual dataset, eliminated some use of an expensive and limited clinical data, but allowed for creating, testing, and demonstrate a working AI based tool.

FUTURE DIRECTIONS

Validation with Real Genomic Data:

The current system was built on a self-created dataset simulating patient characteristics and genomic profiles. The next step would be to integrate real genomic datasets such as IndiGenome in oracle with a standardized notation format similar to the design in this project to confirm clinical relevance and ensure real-world usability.

The Addition of Clinical Annotations:

By adding clinical annotations (diagnosis, drug responses, treatment history) to datasets like IndiGenome, their import for personalized medicine systems will be critical. This would provide end-to-end prediction from genome to treatment.

The Incorporation of More Genes and Drug Mappings:

Future versions of the model can accommodate more genomic markers (e.g., EGFR, KRAS, CYP2D6) and wider drug classes to cover more disease domains (neurology, cardiovascular).

Interoperability with Health Records:

Provided privacy and data standards are kept, the model can extend to assuming interoperability with Electronic Health Records (EHR) to add value for real-time clinical decisioning.

CONCLUSION

This paper proposes a framework approach to AI assisted personalized medicine, with the idea of genomics being the base underlying treatment recommendations. While the original concept was to depend on real-world genomic entities such as IndiGenome, and PharmGKB, it became evident that their format compatibilities and annotations made using these resources too problematic to proceed with as planned in the project.

To establish whether the framework could work in some way, a dataset was manually contrived using standardized notations and gene markers (BRCA1, BRCA2, TP53) for the training of a Random Forest model and deployment via a Streamlit-based application which could take user uploaded files that could be processed to generate treatment recommendations along with dietary and lifestyle recommendations.

Although this implementation should not be construed to be clinically validated; at this stage we have a technical workflow validated along with a basis for us to refine. Once real datasets come into play, and using a compatible structure, both functional (to be placed in practice) and patient centered, our implementation should adapt from this point easily into something useful. All together this project provides a meaningful step toward establishing AI assisted personalized medicine in a way that aspirationally will be accessible, explainable, and scalable in an Indian context.

REFERENCES

[1] Khurana, V., & Mishra, P. (2021). Applications of Artificial Intelligence in Genomics: From Data to Precision Medicine. Journal of Precision Medicine, 18(3), 112-128.

[2] Ritchie, M. D., & Newhouse, S. J. (2018). Pharmacogenomics: Translating the Science of Genomics into Personalized Medicine. Wiley & Sons.

[3] Wang, K., & Li, M. (2020). PharmGKB: A Pharmacogenomic Resource for Predicting Drug Response. Pharmacogenomics Journal, 12(4), 211-220.

[4] Sood, R., & Verma, R. (2022). Personalized Medicine and Genomic Data: The Future of Healthcare in India. Indian Journal of Medical Research, 157(6), 872-879.

[5] Liu, Y., & Xie, Y. (2019). Machine Learning for Personalized Medicine: The Role of Big Data and AI in Healthcare Decision Making. Medical Artificial Intelligence, 8(2), 104-118.

[6] Srinivas, K., & Pradeep, A. (2021). Challenges in Genomic Data Integration: Bridging the Gap Between Genomic Variants and Clinical Outcomes. Bioinformatics Advances, 39(1), 64-75.

[7] Kim, J., & Park, S. (2020). AI and Genomic Data: Revolutionizing Precision Medicine with Machine Learning. Journal of AI in Healthcare, 3(1), 22-39.

[8] Pereira, S., & Stewart, J. (2021). Integrating Population-Specific Genomic Data for Personalized Healthcare: Challenges and Opportunities in India. Journal of Global Health, 15(4), 58-72.

[9] Rosenblum, L., & Miller, D. (2019). Ensemble Learning for Predicting Drug Efficacy in Genomic Medicine. Computational Biology and Chemistry, 76, 42-50.

[10] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems, 30, 4765-4774.

[11] Caruana, R., & Gehrke, J. (2015). Use of Local Surrogate Models to Explain Black-box Classifiers. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 1-10.

[12] Sanyal, A., & Shah, M. (2022). Ethical Considerations in the Use of AI for Genomic Medicine. AI and Ethics, 3(1), 29-41.

[13] Denny, J. C., & Ritchie, M. D. (2021). Leveraging Genomic and Electronic Health Record Data for Precision Medicine. Journal of the American Medical Association (JAMA), 325(18), 1818-1830.

[14] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

[15] Easton, D. F., et al. (2015). Gene-panel sequencing and the prediction of breast-cancer risk. New England Journal of Medicine, 372(23), 2243–2257.

[16] Medical diagnosis by using machine learning and deep learning techniques: A review. Computers in Biology and Medicine, Volume 144, 105284.