Design Thinking Meets Generative AI: Shaping the Future of Medical Diagnostics

Murali Krishnan Jayagopi, MSc in Computing Big Data Analytics and Artificial Intelligence

***Abstract*— Advancements in Artificial intelligence and machine learning are revolutionizing healthcare, especially in predictive diagnostics where early and accurate disease outcome prediction significantly impacts patient care. This study combines ML techniques with user-centered design principles through Design Thinking to develop a disease outcome prediction system. Using a dataset with 132 symptom-based variables and employing ensemble methods such as Random Forest and Hist gradient boosting, we aim to create a high-accuracy diagnostic model that aligns with the real-world needs of healthcare providers. Iterative testing and model refinement, supported by Design Thinking, ensure the diagnostic tool's relevance and usability in clinical settings.**

I. INTRODUCTION

Machine learning and artificial intelligence are rapidly advancing healthcare, with predictive diagnostics playing a critical role in enhancing patient care through early and accurate disease outcome prediction. By combining machine learning (ML) with a user-centered design process—more especially, by utilizing Design Thinking methodologies—this project seeks to create a reliable disease outcome prediction system. In order to create efficient diagnostic models for practical applications, Design Thinking offers an organized, iterative framework that prioritizes comprehending user demands, identifying fundamental problems, coming up with solutions, prototyping, and testing.

The dataset used comprises 132 symptom-based features and an outcome column capturing the diagnosed disease. Ensemble methods, specifically Random Forest and Hist gradient boosting, are employed to handle this high-dimensional data. Ensemble classifiers combine multiple algorithms to enhance predictive accuracy and reduce errors, making them particularly suitable for datasets with numerous features. Through iterative model development and testing, this project aims to achieve not only high diagnostic accuracy but also a solution that meets healthcare providers' needs. This alignment, facilitated by Design Thinking and enhanced by Generative AI tools, has enabled continuous user feedback to refine the model's practical utility in clinical settings.

II. MOTIVATION

The need for precise and rapid diagnostic tools in healthcare drives this project. With vast amounts of patient data and symptom information, healthcare providers face the challenge of identifying patterns relevant to accurate disease diagnosis. Predictive models offer a pathway to simplify this process, providing clinicians with reliable, data-driven insights to aid decision-making. This demand is particularly strong in emergency care, preventive health, and initial patient assessments, where prompt diagnosis improves patient outcomes. Enabling fast and accurate predictions could reduce clinicians' diagnostic workload, facilitating quicker and more accurate treatment decisions.

Our approach is based on Design Thinking principles, emphasizing healthcare practitioners' specific needs. Feedback from the Empathize and Define stages highlighted that practitioners expect the model to integrate seamlessly with existing workflows and deliver reliable, actionable predictions. This insight influenced the decision to employ ensemble classifiers, as these models effectively manage large, high-dimensional datasets and reduce overfitting, crucial for producing reliable predictions from complex symptom data. The selected ensemble methods, Random Forest and Hist gradient boosting, offer complementary benefits: Random Forest is robust to missing data and interpretable, while Hist gradient boosting improves prediction accuracy through sequential error correction. Together, these algorithms form the basis of a high-accuracy, user-centered diagnostic model aimed at improving clinical decision-making efficiency and accuracy.

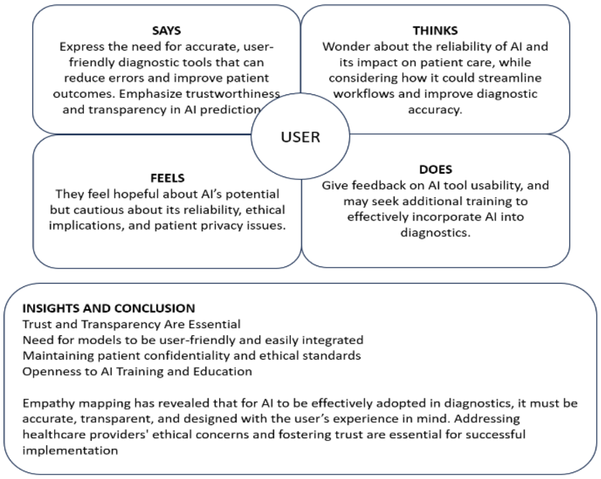
III. DATASET DESCRIPTION AND METHODOLOGY

DATASET OUTLINE

The dataset is divided into two distinct CSV files of 133 columns each for training and testing. In order to assist supervised learning, where symptom data is used to identify likely diseases based on observed data patterns, 132 of these represent symptoms, while one column shows the disease result.

STAGE 1: EMPATHIZE

During the Empathize phase, input was gathered from healthcare experts to align the model's objectives with practical diagnostic needs. An empathy map helped identify key factors:



SAYS
Express the need for accurate, user-friendly diagnostic tools that can reduce errors and improve patient outcomes. Emphasize trustworthiness and transparency in AI prediction

THINKS
Wonder about the reliability of AI and its impact on patient care, while considering how it could streamline workflows and improve diagnostic accuracy.

USER

FEELS
They feel hopeful about AI's potential but cautious about its reliability, ethical implications, and patient privacy issues.

DOES
Give feedback on AI tool usability, and may seek additional training to effectively incorporate AI into diagnostics.

INSIGHTS AND CONCLUSION
Trust and Transparency Are Essential
Need for models to be user-friendly and easily integrated
Maintaining patient confidentiality and ethical standards
Openness to AI Training and Education

Empathy mapping has revealed that for AI to be effectively adopted in diagnostics, it must be accurate, transparent, and designed with the user's experience in mind. Addressing healthcare providers' ethical concerns and fostering trust are essential for successful implementation

Fig 1 Empathy Map

- SAYS: The dataset represents a range of symptoms linked to potential diagnoses.
- THINKS: Healthcare users expect the model to increase diagnostic speed and reliability.
- FEELS: Users acknowledge the dataset's complexity and recognize the model's potential in supporting differential diagnosis.
- DOES: The model should integrate with current diagnostic processes to serve as a useful clinical tool.

STAGE 2: DEFINE

In the Define phase, insights from healthcare professionals clarified key challenges and goals. With numerous symptom variables, managing high-dimensional data and avoiding overfitting became priorities. Ensemble classifiers were chosen for their efficiency in handling multiple input variables and mitigating overfitting risks, with the goal of delivering accurate predictions that meet healthcare professionals' diagnostic requirements.

STAGE 3: IDEATE

In the Ideate phase, various model structures were considered, with ensemble methods emerging as the most suitable. Random Forest and Hist gradient boosting were selected for their ability to handle large datasets with numerous features. Random Forest handles categorical variables effectively and is robust against missing data, while Hist gradient boosting improves prediction accuracy by sequentially correcting previous errors.

STAGE 4: PROTOTYPE

In the Prototype stage, Missing values were managed through imputation or, where necessary, discarding incomplete records. After splitting the data into training and testing sets, ensemble models were constructed:

- Random Forest: Uses random data subsets to train several decision trees; the final result is determined by a majority vote.

- Hist gradient boosting: Sequentially enhances accuracy by correcting prior model errors, with a histogram-based approach that accelerates processing, especially for large datasets.

STAGE 5: TEST

The final stage evaluated model performance using metrics like accuracy, precision, recall, and robustness. Testing Random Forest and Hist gradient boosting on symptom interactions showed that Random Forest provided a strong baseline accuracy, while Hist gradient boosting excelled in managing complex interactions and residual error correction.

IV. ENSEMBLE CLASSIFIERS OVERVIEW

ENSEMBLE CLASSIFIERS

Ensemble classifiers combine predictions from multiple models to improve accuracy and reduce error. This approach is especially effective for high-dimensional datasets, as it balances each model's strengths and weaknesses.

A. RANDOM FOREST

Random Forest, an ensemble technique based on bagging, constructs numerous decision trees on various data subsets and aggregates their predictions. This model is suitable for both categorical variables and missing data, with the bagging approach enhancing Random Forest's resilience against overfitting.

B. HIST GRADIENT BOOSTING

Hist gradient Boosting, a form of Gradient Boosting, refines models sequentially to reduce prior prediction errors. By discretizing continuous features with histograms, it speeds up processing, making it ideal for datasets with intricate symptom-outcome relationships.

V. ETHICAL REFLECTION

The ethical application of AI in medical diagnostics presents challenges, especially related to bias, privacy, and transparency. To mitigate biases, we used a varied dataset and conducted frequent audits of the model's predictions to ensure fairness. Patient data privacy was protected through anonymization and adherence to robust data protection standards. For transparency, we utilized Generative AI to highlight key symptoms influencing each diagnosis, helping users better understand the model’s decision-making process.

VI. CONCLUSION

This study demonstrates the effectiveness of Random Forest and Hist gradient boosting in predicting disease outcomes from symptom data. Random Forest provided reliable base- line accuracy and handled missing data effectively, while Hist gradient boosting improved accuracy through residual correction. Both models achieved a final accuracy of 97.62%, validating their reliability for high-dimensional diagnostic applications. Future work may include further refinement of feature engineering and exploring additional ensemble methods to enhance diagnostic precision.

VII. REFERENCES

1. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
Available:
https://link.springer.com/article/10.1023/A:1010933404324

2. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016.
Available: https://dl.acm.org/doi/10.1145/2939672.2939785

3. T. Brown, "Design Thinking," *Harvard Business Review*, vol. 86, no. 6, pp. 84-92, 2008.
Available: https://hbr.org/2008/06/design-thinking

4. A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, and J. Dean, "A Guide to Deep Learning in Healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24-29, 2019.
Available: https://www.nature.com/articles/s41591-018-0307-4

5. E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44-56, 2019.
Available: https://www.nature.com/articles/s41591-018-0300-y