Project Title

A Project Report

submitted in partial fulfillment of the requirements

of

AICTE Internship on AI: Transformative Learning
with
TechSaksham – A joint CSR initiative of Microsoft & SAP

by

Shrivardhan Bangale

Shrivardhan.bangale@dypic.in

Under the Guidance of

Saomya Chowdary

ACKNOWLEDGEMENT

I would like to take this opportunity to express my sincere gratitude to everyone who has supported and guided me throughout this project. Their valuable insights, encouragement, and contributions have been instrumental in the successful completion of this work.

Firstly, I would like to extend my heartfelt thanks to our supervisor **Saomya Chowdary** for his invaluable guidance, constant support, and expert advice. His encouragement and constructive feedback have been a great source of motivation and learning throughout the project. The confidence and trust shown in me have inspired us to strive for excellence. It has been an absolute privilege to work under his mentorship, and I am truly grateful for the knowledge and skills that I have gained under his supervision.

I would also like to thank **TechSaksham and faculty members** for providing me with the necessary resources, facilities, and academic support that enabled us to carry out this research effectively. Their insightful suggestions and encouragement have significantly contributed to my understanding and execution of the project.

Furthermore, I express my deep appreciation to our **family and friends** for their continuous support, patience, and motivation throughout this journey. Their unwavering belief in me has been a pillar of strength, helping us overcome challenges and stay committed to our goals.

Lastly, I extend my gratitude to all the researchers, authors, and developers whose work and contributions in the field of **AI and medical diagnosis** have provided us with the foundation to build upon.

This project has been a tremendous learning experience, and I am so grateful for the opportunity to apply our knowledge in a practical and impactful way.

ABSTRACT

This project focuses on building a **Disease Prediction System** using **Machine Learning (ML)**. The system predicts multiple diseases, including **Diabetes, Heart Disease, Parkinson's, Lung Cancer, and Hypo-Thyroid conditions** based on user input data. The primary objective is to develop an **AI-driven diagnostic assistant** that helps individuals assess their health conditions quickly and efficiently.

With the increasing prevalence of **chronic and lifestyle-related diseases**, early diagnosis plays a crucial role in improving patient outcomes. However, traditional diagnostic methods require extensive medical examinations, laboratory tests, and expert consultations, which may not always be accessible, especially in remote areas. This project aims to bridge that gap by providing an easy-to-use, **web-based predictive system** that utilizes trained machine-learning models to deliver **instant** predictions based on user input.

The system is implemented using **Streamlit**, a popular Python framework for creating interactive web applications. The backend is powered by **pre-trained ML models** stored using the **pickle** module. These models analyse key **medical parameters** provided by the user and predict the likelihood of specific diseases based on trained datasets. The diseases covered in this project have been selected due to their **high global incidence rates and critical health impact**.

Users are required to input their health parameters, such as **blood glucose levels, heart rate, body mass index (BMI), lifestyle habits, and medical history**. The system processes this data through trained machine-learning models and instantly provides **disease predictions with recommendations**. This application can be **integrated into telemedicine services** to assist in **early disease detection and medical consultation**.

The **accuracy** of predictions depends on the quality of input data and the effectiveness of the trained models. This project demonstrates how **AI in healthcare** can enhance **accessibility, affordability, and early diagnosis**, ultimately contributing to better health outcomes and preventive care strategies.

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

Introduction

1.1Problem Statement:

Early detection of diseases is crucial for effective treatment and improved patient outcomes. However, traditional diagnostic methods often involve lengthy medical checkups, expensive lab tests, and multiple consultations, making timely diagnosis challenging, especially in remote or underserved areas. With the rise in chronic diseases such as diabetes, heart disease, and cancer, there is a growing need for an automated, AI-powered disease prediction system that can analyze patient data instantly and provide accurate preliminary diagnoses. Such a system can serve as a decision-support tool for individuals and healthcare professionals, facilitating early intervention and reducing the burden on healthcare facilities.

1.2Motivation:

Advancements in Artificial Intelligence (AI) and Machine Learning (ML) have revolutionized healthcare by enabling automated, data-driven diagnostics. Traditional disease detection relies on extensive medical tests, which can be expensive, time-consuming, and inaccessible to individuals in rural or underserved regions.

This project is motivated by the need for an AI-powered, user-friendly web application that allows individuals to self-assess their health by entering key medical parameters. By leveraging ML models trained on medical datasets, the system can provide quick and reliable predictions for diseases such as Diabetes, Heart Disease, Parkinson’s, Lung Cancer, and Hypo-Thyroid conditions.

Such a system has immense potential to assist medical professionals by serving as a pre-diagnostic tool, enabling early intervention, and reducing hospital workload. Additionally, it empowers individuals with health awareness, making preventive care more accessible and efficient.

1.3Objective:

- Develop an AI-based Disease Prediction System
- Implement a web-based interface using Streamlit
- Utilize Machine Learning models for accurate predictions
- Provide a user-friendly experience with real-time results
- Enhance accessibility to preliminary medical checkups

1.4Scope of the Project:

- Covers five diseases: Diabetes, Heart Disease, Parkinson's, Lung Cancer, and Hypo-Thyroid
- Uses trained ML models to make predictions
- Can be expanded to include more diseases in the future
- Can be integrated with telemedicine platforms

Limitations

- Dependence on Input Data Quality
- Lack of Real-Time Medical Data Integration
- Security and Privacy Concerns
- Bias in Machine Learning Models

CHAPTER 2

Literature Survey

2.1 Review relevant literature or previous work in this domain.

The integration of Artificial Intelligence (AI) and Machine Learning (ML) in healthcare has been a growing field of research, with numerous studies demonstrating their effectiveness in disease prediction and diagnosis. Several researchers have explored the use of machine learning algorithms to analyze medical datasets and predict diseases with high accuracy. For instance, in diabetes prediction, studies have shown that Logistic Regression, Decision Trees, Random Forest, and Neural Networks achieve promising results when trained on datasets such as the PIMA Indian Diabetes Dataset. Similarly, research in cardiovascular disease prediction highlights the effectiveness of Support Vector Machines (SVM), Naïve Bayes, and Deep Learning models in detecting heart-related conditions based on key physiological indicators. Additionally, Parkinson’s disease prediction has been explored using speech signal analysis and neuroimaging data, with K-Nearest Neighbors (KNN), Deep Neural Networks (DNN), and Decision Trees demonstrating good predictive performance. Studies on lung cancer detection have incorporated radiographic imaging analysis with Convolutional Neural Networks (CNNs), while thyroid disease classification has utilized Gradient Boosting Machines and XGBoost.

2.2 Mention any existing models, techniques, or methodologies related to the problem.

Various machine learning techniques have been implemented for disease prediction:

1) **Logistic Regression (LR)** – Commonly used for binary classification, especially in predicting diabetes and heart disease.
2)**Decision Trees (DT)** – Effective in rule-based disease classification but prone to overfitting.
3)**Random Forest (RF)** – An ensemble learning method that improves accuracy by reducing overfitting.
4)**Support Vector Machine (SVM)** – Used for high-dimensional datasets, particularly in Parkinson’s disease detection.
5)**K-Nearest Neighbors (KNN**) – Applied in medical image analysis and disease classification.
6)**Neural Networks (NN)** – Deep Learning models that excel in complex feature extraction, particularly in cancer detection.
7)**Convolutional Neural Networks (CNNs)** – Used in medical imaging for diagnosing lung cancer and other radiological diseases.

These methodologies have significantly improved diagnostic efficiency, but they also come with challenges that limit their real-world applicability.

2.3 Highlight the gaps or limitations in existing solutions and how your project will address them.

While AI-driven medical diagnostics have shown great potential, existing solutions have notable limitations:

- **Data Quality and Availability:** Most ML models require extensive high-quality, labeled medical datasets, which may not always be available. Public datasets are often limited in size and diversity.

- **Generalization Issues:** Some AI models perform well on specific datasets but fail to generalize to real-world patient data, limiting their clinical adoption.

- **Interpretability and Explainability:** Many deep learning-based models act as black boxes, making it difficult for healthcare professionals to interpret predictions and trust AI-generated results.
- **Computational Requirements:** Some advanced models, such as CNNs and Deep Neural Networks, require high computational power, making them difficult to deploy in resource-limited settings.
- **Integration with Healthcare Systems:** Many existing AI models are developed in isolation and are not seamlessly integrated into real-world hospital databases and telemedicine platforms.

2.4 How Our Project Addresses These Gaps

This project aims to overcome these challenges by:

- **Using Pre-trained Machine Learning Models:** By utilizing trained models with pickle, our system reduces computational overhead, making it lightweight and accessible.
- **Providing a Web-Based Interface with Streamlit:** Unlike existing standalone AI models, our system is user-friendly and can be easily accessed via a web browser, ensuring accessibility even in remote areas.
- **Focusing on Multiple Diseases:** Instead of predicting a single disease, our project integrates multiple disease detection capabilities (Diabetes, Heart Disease, Parkinson's, Lung Cancer, and Hypo-Thyroid) into one system, increasing its utility.
- **Enhancing Interpretability:** By displaying user-friendly results and guiding users with health recommendations, our system makes AI predictions more understandable and actionable.
- **Scalability and Future Expansion:** The framework can be extended to include additional diseases by training new models, making it adaptable to future advancements in medical AI.

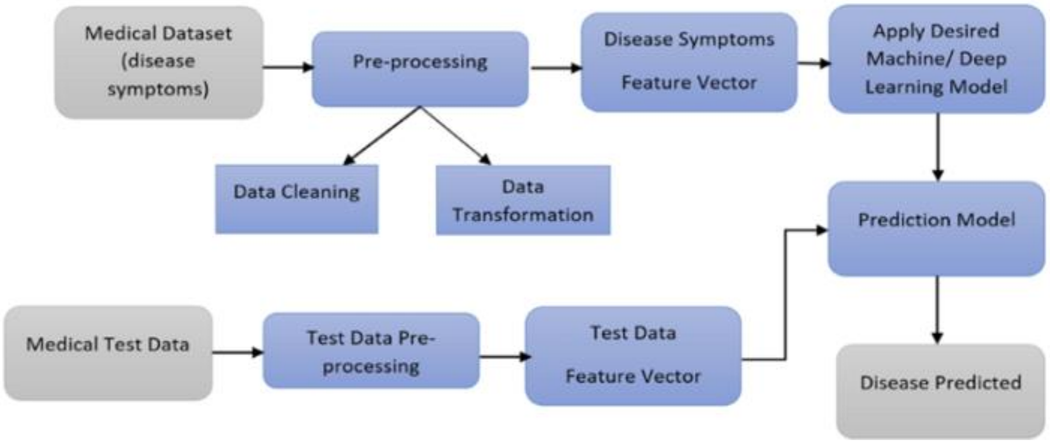
By addressing these gaps, our **Disease Prediction System** contributes to the growing field of AI-assisted diagnostics, offering a **practical, accessible, and scalable solution** for early disease detection and preventive healthcare.

CHAPTER 3

Proposed Methodology

3.1 System Design

System Architecture



The workflow of an **AI-powered medical diagnosis system**. It outlines the key steps involved in processing medical data, applying machine learning models, and predicting diseases. Below is a breakdown of the process:

1. Medical Dataset (Disease Symptoms)

- The system begins with a **medical dataset** that contains patient records, symptoms, and diagnostic results.
- This dataset serves as the foundation for training machine learning models.

2. Pre-processing

- The raw dataset undergoes **pre-processing**, which includes:
  - **Data Cleaning**: Handling missing values, removing duplicates, and correcting inconsistencies.
  - **Data Transformation**: Converting raw data into a structured format suitable for machine learning models.

- After pre-processing, the data is transformed into **feature vectors**, which represent disease symptoms in numerical form.

3. Machine Learning Model Application

- The **feature vector** is fed into a **Machine Learning (ML) or Deep Learning (DL) model**.
- Various models can be used, such as **Decision Trees, Support Vector Machines (SVM), Random Forest, or Neural Networks**.
- The selected model processes the input data and generates predictions.

4. Prediction Model

- The trained model takes in **test data** (new patient records) and processes it using the same **pre-processing pipeline**.
- The test data is transformed into a **feature vector** similar to the training data.

5. Disease Prediction

- The trained prediction model outputs a result, indicating the likelihood of a disease based on the input data.
- The final output is a **diagnosis prediction**, which can be used for early detection and further medical consultation.

Key Takeaways

- The system follows a **structured pipeline** that ensures efficient disease prediction.
- **Pre-processing** ensures data quality before model training.
- **Machine learning models** play a crucial role in making accurate predictions.
- The system can be integrated into **telemedicine applications** for **real-time diagnosis**.

3.2 Requirement Specification

3.2.1 Hardware Requirements:

To efficiently run the application and perform machine learning-based disease predictions, the following hardware specifications are recommended:

- **RAM**: Minimum 4GB (Higher RAM is recommended for faster processing)
- **Processor:** Intel i5 or above (or equivalent AMD processor)
- **Storage:** At least 10GB of free disk space (to store models and dependencies)
- **Internet Connectivity:** Required for deploying and using the web-based interface

A system with higher specifications, such as 8GB+ RAM and an SSD, is recommended for faster model execution and better user experience.

3.2.2 Software Requirements:

The system relies on several software components for data processing, model execution, and web-based deployment:

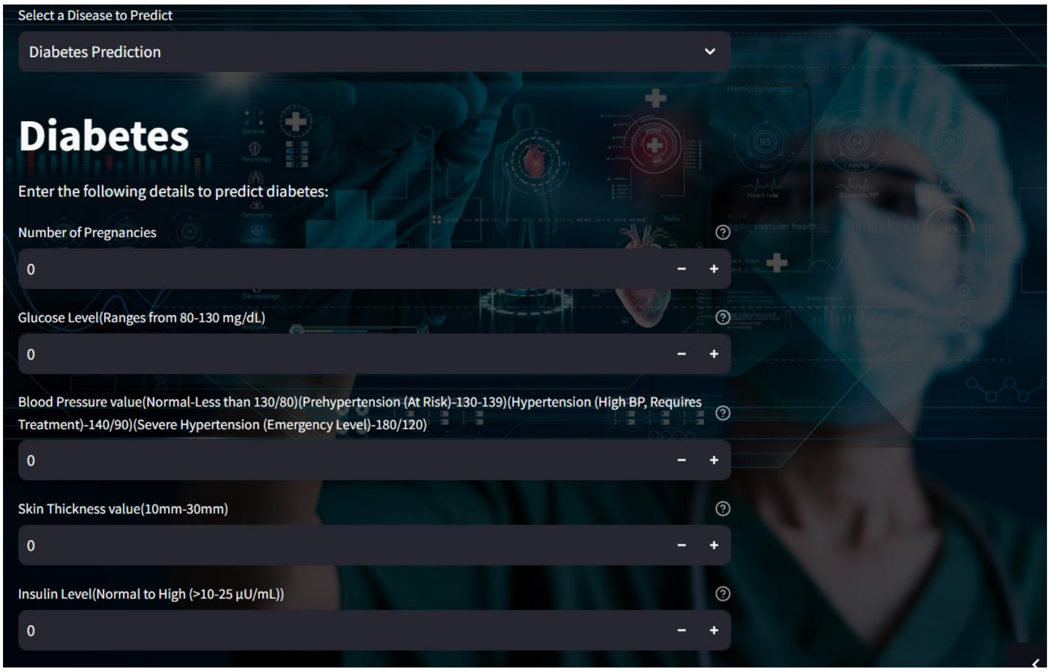
- **Operating System**: Windows, Linux, or macOS
- **Programming Language**: Python (**v3.x**)
- **Frameworks & Libraries**:
  - **Streamlit** – For developing the interactive web application
  - **Scikit-learn** – For machine learning model training and prediction
  - **Pickle** – For saving and loading trained ML models
  - **Pandas & NumPy** – For handling and processing data
  - **Matplotlib & Seaborn** – For data visualization (optional)

The **application can be hosted on cloud platforms** like **Streamlit Cloud, AWS, or Heroku** for easy accessibility.

CHAPTER 4

Implementation and Result

4.1 Snap Shots of Result:



Select a Disease to Predict

Diabetes Prediction

Diabetes

Enter the following details to predict diabetes:

Number of Pregnancies

0

Glucose Level(Ranges from 80-130 mg/dL)

0

Blood Pressure value(Normal-Less than 130/80)(Prehypertension (At Risk)-130-139)(Hypertension (High BP, Requires Treatment)-140/90)(Severe Hypertension (Emergency Level)-180/120)

0

Skin Thickness value(10mm-30mm)

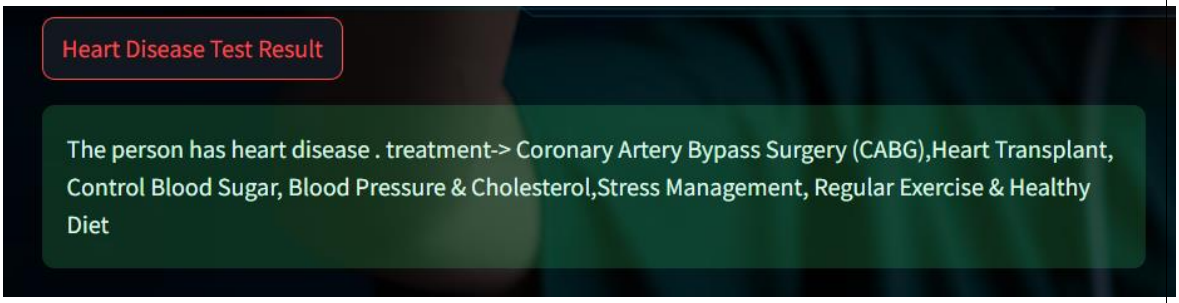
0

Insulin Level(Normal to High (>10-25 µU/mL))

0

Fig 1. Home Page

In this Figure, the user is welcomed with the Model and its Interactive UI. Upon this He/She has to enter the data as per the Column Field.

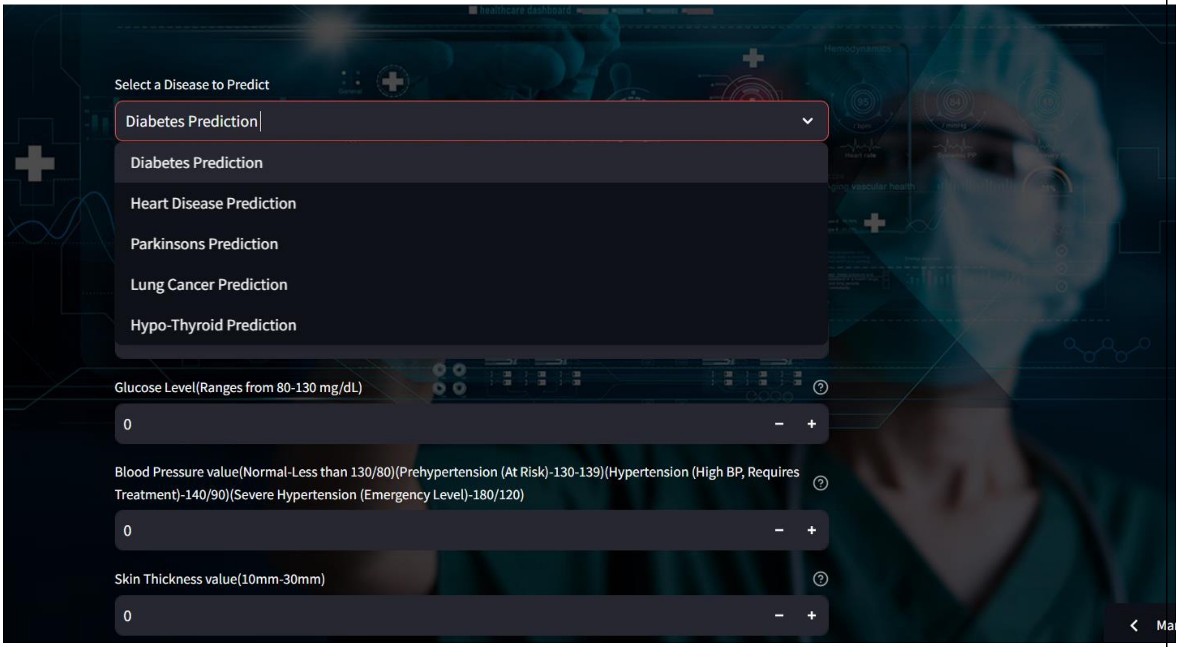


Heart Disease Test Result

The person has heart disease . treatment-> Coronary Artery Bypass Surgery (CABG),Heart Transplant, Control Blood Sugar, Blood Pressure & Cholesterol,Stress Management, Regular Exercise & Healthy Diet

Fig 2. Output/Result

In this Figure, after entering the required Data the models Analyzes all the Fields and Predicts the Result for the given Disease. In the Above Figure, the Patient is Diagnosed with Heart disease and the Treatment is also listed for it.



Select a Disease to Predict

Diabetes Prediction

Glucose Level(Ranges from 80-130 mg/dL)

0

Blood Pressure value(Normal-Less than 130/80)(Prehypertension (At Risk)-130-139)(Hypertension (High BP, Requires Treatment)-140/90)(Severe Hypertension (Emergency Level)-180/120)

0

Skin Thickness value(10mm-30mm)

0

Fig 3. List of choices

In the Above Figure, The User can choose between any of the 5 Listed Disease Prediction Models and use it according to them.

4.2GitHub Link for Code:

https://github.com/shrivardhanBangale16/Medical-Diagnosis-Using-AI-ML

CHAPTER 5

Discussion and Conclusion

5.1 Future Work:

The current Disease Prediction System provides a robust foundation for AI-driven medical diagnostics. However, there are several areas for improvement and expansion:

5.1.1 Extend the System to Predict More Diseases

- Currently, the system supports diseases like Diabetes, Heart Disease, Parkinson’s, Lung Cancer, and Hypo-Thyroid.
- Future versions can incorporate predictions for more diseases such as Alzheimer’s, Kidney Disease, Hypertension, and Stroke.
- Expanding the dataset with diverse medical conditions will enhance the system’s usability.
- Expanding the dataset with diverse medical conditions will enhance the system’s usability.

5.1.2 Improve the Accuracy of ML Models

- Enhance prediction accuracy by using more advanced deep learning techniques such as Neural Networks and Ensemble Learning.
- Implement hyperparameter tuning and optimize models using GridSearchCV or Bayesian Optimization.
- Increase training data size by integrating larger and more diverse datasets.

5.1.3 Integrate with Hospital Databases for Better Diagnosis

- Connect the system with electronic health records (EHRs) and hospital databases to provide real-time data-driven diagnosis.
- Enable doctors to review AI predictions alongside actual medical records for better decision-making.
- Incorporate IoT-based health monitoring devices for real-time patient tracking and automatic parameter updates.

5.2 Conclusion:

The AI-powered Disease Prediction System successfully leverages Machine Learning and Streamlit to provide an interactive and user-friendly medical diagnosis platform. The project achieves the goal of predicting multiple diseases based on user inputs, enabling early detection and preventive healthcare.

By integrating this tool with telemedicine platforms, users can receive quick health assessments without requiring frequent hospital visits. Future enhancements, such as increasing prediction accuracy, adding more diseases, and integrating with real-world medical databases, will further improve the system's reliability and impact.

REFERENCES

[1]. Ming-Hsuan Yang, David J. Kriegman, Narendra Ahuja, “Detecting Faces in Images: A Survey”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume. 24, No. 1, 2002.

[2]. **Patel, S., Park, H., Bonato, P., Chan, L., & Rodgers, M. (2012).** “A review of wearable sensors and systems with application in rehabilitation.” Journal of NeuroEngineering and Rehabilitation, 9(1), 21.

[3]. **Chicco, D., & Jurman, G. (2020).** "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone." BMC Medical Informatics and Decision Making, 20(1), 1-16.

[4]. **Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., & Ng, A. Y. (2017).** "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning." arXiv preprint arXiv:1711.05225.

[5]. **Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017).** "Dermatologist-level classification of skin cancer with deep neural networks." Nature, 542(7639), 115-118.

[6]. **Tang, Y., & Huber, M. (2017).** "Automatic detection of Parkinson’s disease using machine learning methods." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[7]. **Gao, X. W., Hui, R., & Tian, Z. (2017).** "Classification of CT brain images based on deep learning networks." Computer Methods and Programs in Biomedicine, 138, 49-56.

[8]. **Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017).** "Artificial intelligence in precision cardiovascular medicine." *Journal of the American College of Cardiology.