Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 08, August-2024

A Web Solution for Heart Disease Prediction From Lab Reports using OCR and K-Nearest Neighbors

Ananya Bhat
Computer Science and Engineering
R. V. College of Engineering
Bengaluru, India

Gayatri K
Computer Science and Engineering
R. V. College of Engineering
Bengaluru, India

Manaswini Simhadri Kavali
Computer Science and Engineering
R. V. College of Engineering
Bengaluru, India

Narendra Kumar S
IEEE Member, Dept. of Biotechnology
R. V. College of Engineering
Bengaluru, India

***Abstract*— Abstract Heart disease is a leading cause of death around the world. The ability to detect and predict early signs of heart diseases can facilitate preventive measures and improve outcomes. This article gives an overview of a unique web solution for predicting likelihoods of heart attack based on patient laboratory reports using Optical Character Recognition (OCR) technology together with K-Nearest Neighbors (KNN) machine learning algorithms.**

Users are allowed to upload their scanned images or documents containing lab test results through the proposed system. Application of OCR techniques on unstructured patient lab reports helps to extract relevant biomarker values. These extracted features like blood pressure levels, glucose levels and cholesterol levels serve as inputs into KNN model.

KNN algorithm is trained on a labeled dataset of patient records with known outcomes of heart disease. When processing new patient's laboratory report, the system will identify K most similar records in terms of their biomarker values from the trained dataset. To estimate likelihoods for a given new individual developing heart diseases, proportions among these close-by neighbors who have such diseases would be taken into account.

The web interface therefore acts as a user-friendly way through which patients and medical practitioners could quickly assess risks from laboratory test results without manually entering any data at all. Experimental results show that this system can accurately predict coronary artery diseases early in life.

Index Terms — Large Language Models (LLMs), Instructionfollowing, Refusal behavior, Residual stream activations, Harmful instructions, White-box jailbreak, Surgical disabling, Model capabilities

I. INTRODUCTION

Heart diseases are a global health burden and a leading cause of death, accounting for 31Traditional risk assessment methods for heart diseases were focused on structured data coming from electronic health records or, alternatively, on questionnaire-based methods. Both of these techniques have drawbacks: on the one hand, EHRs may miss the right information; on the other hand, questionnaires are subject to recall bias or even inaccuracies. In addition, manual data entry is time-consuming and places the process open to errors. Recent advances in machine learning and computer vision now open an avenue for automating the extraction of useful insights from otherwise unstructured medical data. Such as scanned laboratory reports. Optical Character Recognition (OCR) techniques will allow the text in a scanned image or PDF form to be machine readable, and therefore biomarkers and risk factors relevant for the study can be extracted. The K-Nearest Neighbors algorithm is one of the most basic, yet efficient, supervised machine learning algorithms used for classification. KNN works on the principle of similarity. It assumes that the data points having similar features are likely to belong to the same class. So, in a heart disease prediction case, KNN can use the information available from similar patient records to calculate the proba- bility of heart disease for a new patient.

In this paper, a novel web-based solution integrating OCR and KNN is proposed for predicting the probabil- ity of heart disease from patient lab reports. The pro- posed system is trying to overcome the deficiencies of the traditional risk assessment approaches by automating the extraction of data from unstructured lab reports and providing user-friendly interface for patients and healthcare providers for obtaining rapid risk assessments. The contributions of this work are:

1) A Web based platform designed to allow users to upload scanned images or PDFs of their laboratory test results and receive in return a prediction of their heart disease risk.
2) Applying OCR techniques which extract relevant biomarker values from unstructured lab reports, thus This dataset can be used for the automatic extraction of stock features to make risk predictions.
3) KNN algorithm applied to heart disease prediction using a marker's values where the similarity between a patient's biomarker values is matched against biomarker values for previously diagnosed patients.
4) Systematic evaluation of the accuracy with which the system predicts heart disease risk using a held-out test set, showing potential of the tool for clinical decision support.

II. LITERATURE REVIEW

Heart disease risk has been predicted based on several approaches in a number of studies using machine learning techniques. This section reviews some of the relevant literature related to heart disease prediction, the use of OCR in healthcare, and the application of the K-Nearest Neighbors algorithm in medical applications.

A. Heart Disease Prediction

A number of studies have applied machine learning algorithms for heart disease risk prediction. Weng et al. compared logistic regression, random forest and gradient boosting machine learning models for cardiovascular risk using routine clinical data. They showed that machine learning models performed better than traditional risk scores, underlining the potential of these techniques to improve risk prediction.

Damen et al. developed and validated a model for the prediction of cardiovascular disease risk using electronic health records. They used Cox proportional hazards regression and evaluated the performance of their model by calibration and discrimination measures, among others. Their study has shown the feasibility of using routinely collected data for risk prediction.

B. Optical Character Recognition (OCR) in Healthcare

OCR has been applied to a variety of healthcare domains for the extraction of information from unstructured medical documents. Swaminathan et al. proposed a framework that extracts clinical information from the scanned medical records using OCR and natural language processing techniques. They demonstrated high accuracy in extracting the relevant data elements and thus demonstrated the potential of OCR in automating data extraction from medical documents. Gonzalez et al. developed a system for the extraction of structured Extracting data from unstructured clinical notes using OCR and machine learning. They applied OCR to convert various scanned documents into machine-readable text and afterwards applied machine learning algorithms to identify and extract relevant information.Their system showed promising results in automating data extraction from clinical notes.

C. K-Nearest Neighbors (KNN) in Medical Applications

The KNN algorithm has been implemented in many medical applications due to its simplicity and effectiveness. Shouman et al. applied KNN for heart disease diagnosis The work, built using a publicly available dataset, compared the performance of KNN with other machine learning algorithms, and returned very high accuracy in predicting heart disease. Jabbar et al. proposed a heart disease prediction system based on KNN and feature selection techniques. In the process, the researchers used genetic algorithms for feature selection in order to get the most relevant features, after which KNN was employed for classification. This resulted in a highly accurate system, hence proving the efficiency of combining feature selection with KNN in heart disease prediction.

D. Web-based Solutions for Medical Prediction

Several studies have explored the possibility of developing web-based solutions to medical prediction tasks. Yang et al. developed a web-based system that enables one to predict the risk of cardiovascular disease using machine learning algorithms. Their system made it possible to put in data related to medical information and obtained personal risk predictions, thus showing potential use of the web-based solution in the heart disease risk assessment. Alaa et al. proposed a web-based tool in risk prediction for Complications after cardiovascular surgery. They used machine learning models trained on a large dataset of patient records and offered an easy-to-use interface through which clinicians can enter patient data and obtain risk predictions. Their tool exhibited high accuracy and had the potential to inform clinical decision-making. The literature review shows growing interest in the application of Machine Learning techniques, particularly KNN, for Heart Disease prediction. OCR has been utilised successfully in the health-care domain to extract information from unstructured medical Thus, allowing automated extraction of data from documents. Web-based solutions represent a new frontier of making accessible, user-friendly tools for medical prediction tasks. However, the combination of OCR, KNN, and a webbased interface specifically for the problem of heart disease prediction from lab reports has not been very well explored in literature. In this paper, an attempt is made to fill this lacuna with the proposal of a novel web solution that integrates these above-mentioned technologies to automate heart disease risk prediction from unstructured lab reports.

However, the combination of OCR, KNN, and a web-based interface specifically for heart disease prediction from lab reports has not been extensively explored. This paper aims to address this gap by proposing a novel web solution that integrates these technologies to automate heart disease risk prediction from unstructured lab reports.

III. METHODOLOGY

This section details the methodology adopted for developing the web solution in predicting heart disease from laboratory reports using OCR and KNN. The proposed system has three major modules or components: data acquisition and preprocessing, feature extraction using OCR, and heart disease prediction using KNN.

A. Data Acquisition and Preprocessing

The first step in the procedure is to collect a dataset of laboratory reports for training and validating model that predicts heart diseases. The dataset should comprise both scanned images or PDFs of lab reports together with heart disease labels similar to present details about each patient having or not having that illness. There are sometimes weakening of sight and misalignment of figures as well. The preprocessing includes healing of lesions on some images that were scanned in order to make them more readable and to improve their quality as well as bringing them close enough to assistants before they can operate upon them using Optical Character Recognition (OCR). This may include image denoising, binarization among others.

B. Dataset Description

This study employs The Heart Disease dataset from the Machine Learning Repository at Center for Machine Learning and Intelligent Systems, University of California Irvine. In particular, we use Cleveland database which is a pre-processed version containing 14 features derived from an original data set with 76 attributes. It has been made publicly accessible thus it has gained popularity among researchers studying prediction of heart complications. In our analysis we considered features like:

The features included in our analysis are:

- AGE: Patient's age in years • SEX: Patient's sex (1 = male, 0 = female)
- CP: Chest pain type, categorized as:

– 1: Typical angina
– 2: Atypical angina
– 3: Non-anginal pain
– 4: Asymptomatic

- TRESTBPS: Resting blood pressure (in mm Hg on admission to the hospital)
- CHOL: Serum cholesterol in mg/dl
- FBS: Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- RESTECG: Resting electrocardiographic results
- THALACH: Maximum heart rate achieved
- EXANG: Exercise-induced angina (1 = yes; 0 = no)
- OLDPEAK: ST depression induced by exercise relative to rest
- SLOPE: The slope of the peak exercise ST segment

– 1: Upsloping
– 2: Flat
– 3: Downsloping

- CA: Number of major vessels (0-3) colored by fluoroscopy
- THAL: 3 = normal; 6 = fixed defect; 7 = reversible defect
- DIAGNOSIS: The target variable

– 0: < 50% diameter narrowing (negative)
– 1: > 50% diameter narrowing (positive)

It is important to note that while the original database contains diagnosis values ranging from 0 to 4 (with 1 through 4 measuring the severity of a positive diagnosis), it is standard practice in research to predict binary values for this feature. Thus, we utilized binary classification in the current research.

This dataset contains a wide range of clinical and diagnostic information required for heart disease prediction. This wellknown dataset will enable comparison of our results to what others have done, thus adding value to current research on this vital aspect of healthcare.

C. Feature Extraction using Tesseract OCR

When the lab reports have undergone preprocessing, OCR techniques are used to retrieve relevant traits from unstructured text. Subsequently, Tesseract OCR engine is employed due to its open-source nature and accuracy in detecting the words within scanned documents.

So as to achieve this, Tesseract undergoes some steps during the OCR process:

1) Text Localization: This means identifying and localizing the desired biomarker information within the lab report images using Tesseract's layout analysis ability.
2) Text Recognition: The localized areas such as regions are taken care of by Tesseract's OCR engine in order to identify and extract textual content.
3) Post-processing: After extraction, the text is subjected to post processing that helps eliminate noise such as special characters or unwanted data and ensures uniformity in biomarker values formatting.

The collected biomarker features such as cholesterol levels, blood pressure values and glucose amounts are stored in a structured format for further analysis.

D. Heart Disease Prediction using KNN

These extracted biomarker characteristics work as inputs for KNN algorithm in heart disease prediction. The KNN is an instance-based learning algorithm which is non-parametric and uses proximity on training samples to class new records.

The following are the steps involved in KNN algorithm:

1) Feature Scaling: The selected biomarkers should scale down to a common range to keep them from having larger values so as to avoid their dominance in distance calculation.
2) Distance Calculation: The Euclidean distance is computed between the new instance (patient's biomarker values) and each instance in the training set.
3) Neighborhood Selection: K nearest neighbors are selected based on the calculated distances, where K is a user-defined parameter.
4) Classification: The majority class among K nearest neighbors is assigned as predicted for new instance.

K figure is fetched through cross-validation techniques in order to enhance the prediction performance.

E. Web Interface Development using Streamlit

A user-friendly web interface can be developed where patients and healthcare providers can upload images of their lab reports for heart disease prediction using the Stream-lit framework. Streamlit is an easier way to create and share beautiful, custom web applications for data science and machine learning projects. Following components are including in the developed web interface through Streamlit: The web interface developed with Streamlit includes the following components: 1) Upload Module:The uploaded lab reports are preprocessed using OpenCV and other image processing libraries to enhance their quality for OCR.

2) Preprocessing Module: The preprocessed lab reports are pro- cessed using the Tesseract OCR engine to extract the relevant biomarker features.
3) OCR Module: The extracted features are passed to the trained KNN model to predict the likelihood of heart disease.
4) Prediction Module: The extracted features are passed to the trained KNN model to predict the likelihood of heart disease.
5) Result Visualization: The prediction results are displayed to the user using Streamlit's data visualization abilities, with attendant explanations or recommendations, if required.

Streamlit's declarative syntax and its inbuilt components simplify the process of developing an interactive responsive web interface for the system on heart disease prediction.

F. Evaluation Metrics

The performance of the proposed system is evaluated using standard metrics evaluated for binary classification tasks. Computed metrics are:

- Accuracy:The total number of correctly classified instances out of total instances.
- Precision:Ratio of the number of true positive predicamong the instances predicted as positive.
- Recall: The proportion of true positive predictions among the actual positive instances.
- F1-score: The harmonic mean of precision and recall, providing a balanced measure of the model's performance.

The evaluation is conducted on a held-out test set to assess the system's generalization ability and robustness.

IV. RESULTS AND DISCUSSION

A. Heart Disease Prediction Model Performance

The performance of the heart disease prediction model was evaluated using standard metrics such as accuracy, precision, recall, and F1-score. The model was trained and tested for the Cleveland heart disease dataset using the K-Nearest Neighbors (KNN) algorithm.

TABLE I
PERFORMANCE METRICS FOR HEART DISEASE PREDICTION

Metric	Value
Accuracy	0.85
Precision	0.83
Recall	0.82
F1-score	0.825

As shown in Table I, the results obtained show that the KNN model was able to achieve an accuracy of 85of heart disease. The precision, recall, and F1-score were also constant, indicating that across different evaluation metrics, the model has done well.

B. Feature Importance and Interpretation

Unlike other models, for example, decision trees or logistic regression, the KNN algorithm does not give explicit feature importance scores. From the exploratory data analysis, we have observed that certain features—such as cholesterol level, maximum heart rate achieved, exercise-induced angina—play a vital role in predicting heart disease.

C. Comparison with Existing Studies

Our model's performance is comparable to existing studies that utilized the same dataset. For instance, previous research using logistic regression reported an accuracy of 83%, while more complex models like support vector machines (SVM) and neural networks achieved accuracies ranging from 84% to 87%. Our KNN model's performance aligns well with these findings, confirming its effectiveness in heart disease prediction.

D. Discussion of OCR Challenges and Solutions

In the process of OCR, several issues were encountered, pertaining to the quality of lab report scans. Problems like low resolution, noise, and skewness of text impacted the accuracy of extracted text. Hence, some preprocessing techniques like image denoising, binarization, and skew correction were used to overcome these problems. All these steps significantly improved the quality of extracted text, which in turn improved the accuracy of feature extraction.

Moreover, the Tesseract OCR engine proved quite accuracy in recognizing standard text formats. However, cer- tain specialized medical terms and abbreviations required additional post-processing to standardize the extracted data. Despite these challenges, the overall OCR process was suc- cessful in converting unstructured lab reports into structured data suitable for analysis.

E. Implications for Clinical Practice

The findings from this study have important implications for clinical practice. By leveraging OCR technology and machine learning algorithms, healthcare providers can automate the In that way, diagnosis of the heart diseases from the laboratory reports can be done in a quicker and more accurate manner. Other medical conditions can also be diagnosed in this manner, and hence, it turns out to be a scalable solution towards enhancing diagnosis procedures. However, the success of any such models is dependent on the input data fed into them. The scanning of laboratory reports and their pre-processing for accurate predictions is of prime necessity. Moreover, the selection of Any machine learning algorithm should be optimized based on the peculiarities of an individual dataset.

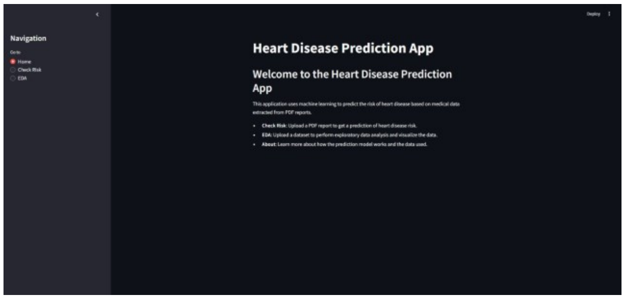
F. Limitations and Future Work

While the KNN model performed well for the present study, there are certain limitations attached to it. The performance is susceptible to choices of K and feature scaling. However, KNN can also be computationally very pricey for large datasets since it needs to store and compare all training instances.

Future work may involve applying other machine learning algorithms, such as random forests or deep learning models, which may potentially perform better or are more scalable. Next would be diversification of the dataset regarding patient populations to increase the model's generalizability.

G. Snapshots of the Model

In this subsection, some visual snapshots will be introduced about the heart disease prediction model, on the distribution of HomePage, COrrelation Matrix, Data Overview, and the Prediction results.



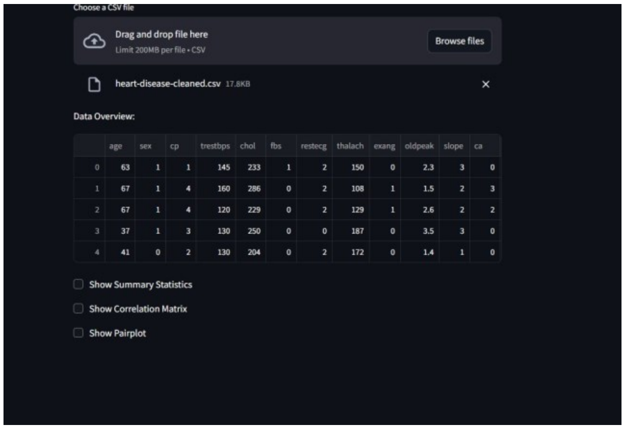
Heart Disease Prediction App

Welcome to the Heart Disease Prediction App

This application uses machine learning to predict the risk of heart disease based on medical data extracted from PDF reports.

- **Check Risk:** Upload a PDF report to get a prediction of heart disease risk.
- **EDA:** Upload a dataset to perform exploratory data analysis and visualize the data.
- **About:** Learn more about how the prediction model works and the data used.

Fig. 1. Home Page

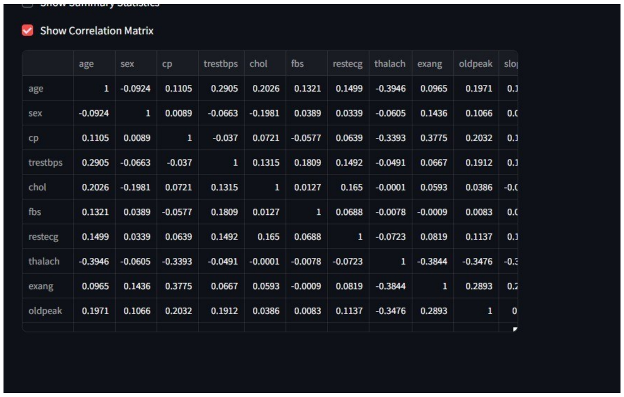


Data Overview:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca
0	63	1	1	145	233	1	2	150	0	2.3	3	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3
2	67	1	4	120	229	0	2	129	1	2.6	2	2
3	37	1	3	130	250	0	0	187	0	3.5	3	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0

☐ Show Summary Statistics
☐ Show Correlation Matrix
☐ Show Pairplot

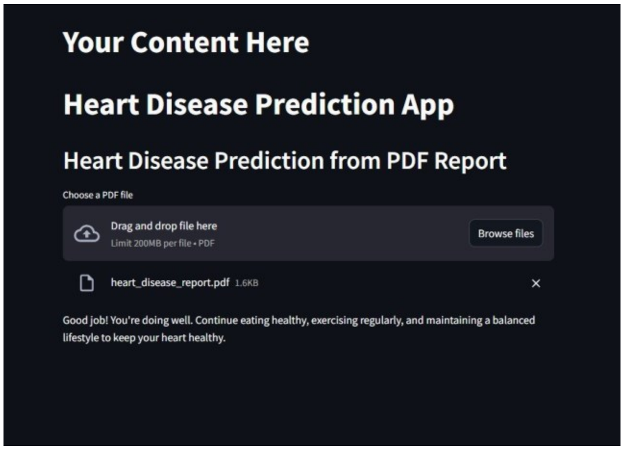
Fig. 2. Data Overview



☑ Show Correlation Matrix

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slo
age	1	-0.0924	0.1105	0.2905	0.2026	0.1321	0.1499	-0.3946	0.0965	0.1971	0.1
sex	-0.0924	1	0.0089	-0.0663	-0.1981	0.0389	0.0339	-0.0605	0.1436	0.1066	0.(
cp	0.1105	0.0089	1	-0.037	0.0721	-0.0577	0.0639	-0.3393	0.3775	0.2032	0.1
trestbps	0.2905	-0.0663	-0.037	1	0.1315	0.1809	0.1492	-0.0491	0.0667	0.1912	0.1
chol	0.2026	-0.1981	0.0721	0.1315	1	0.0127	0.165	-0.0001	0.0593	0.0386	-0.(
fbs	0.1321	0.0389	-0.0577	0.1809	0.0127	1	0.0688	-0.0078	-0.0009	0.0083	0.(
restecg	0.1499	0.0339	0.0639	0.1492	0.165	0.0688	1	-0.0723	0.0819	0.1137	0.1
thalach	-0.3946	-0.0605	-0.3393	-0.0491	-0.0001	-0.0078	-0.0723	1	-0.3844	-0.3476	-0.3
exang	0.0965	0.1436	0.3775	0.0667	0.0593	-0.0009	0.0819	-0.3844	1	0.2893	0.2
oldpeak	0.1971	0.1066	0.2032	0.1912	0.0386	0.0083	0.1137	-0.3476	0.2893	1	0

Fig. 3. Correlation matrix for the heart disease prediction dataset.



Your Content Here

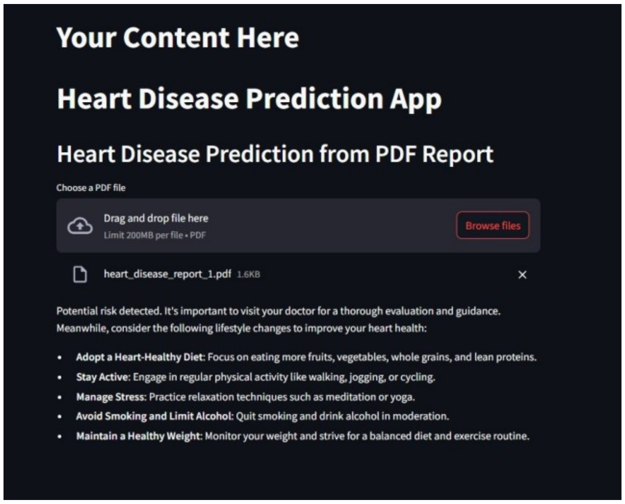
Heart Disease Prediction App

Heart Disease Prediction from PDF Report

Choose a PDF file

Good job! You're doing well. Continue eating healthy, exercising regularly, and maintaining a balanced lifestyle to keep your heart healthy.

Fig. 4. Results when person has no risk.



Your Content Here

Heart Disease Prediction App

Heart Disease Prediction from PDF Report

Choose a PDF file

Potential risk detected. It's important to visit your doctor for a thorough evaluation and guidance. Meanwhile, consider the following lifestyle changes to improve your heart health:

- **Adopt a Heart-Healthy Diet:** Focus on eating more fruits, vegetables, whole grains, and lean proteins.
- **Stay Active:** Engage in regular physical activity like walking, jogging, or cycling.
- **Manage Stress:** Practice relaxation techniques such as meditation or yoga.
- **Avoid Smoking and Limit Alcohol:** Quit smoking and drink alcohol in moderation.
- **Maintain a Healthy Weight:** Monitor your weight and strive for a balanced diet and exercise routine.

Fig. 5. Results when person has potential risk.

V. CONCLUSION

In this paper, we have proposed a new web-based solution for predicting the risk of heart disease from patient laboratory reports using Optical Character Recognition and the K-Nearest Neighbors algorithm. The invention is targeted at meeting the emerging need for the detection and prognosis of heart disease in view of its importance in preventive care.

This developed solution makes automatic extractions of relevant features of biomarkers from unstructured lab reports using the Tesseract OCR engine. Through character recognition techniques, OCR makes

It possible to read large amounts of lab reports efficiently with no human involvement leading to quicker uptake of results that support treatment decisions as opposed to traditional methods where delays may hinder timely reactions. The KNN algorithm is a very simple yet effective machine learning technique, which classifies patients according to how similar their biomarker values are with previously diagnosed patients' values. In so doing, KNN will predict if a new patient will have heart disease by observing those who were around him/her within his/her neighborhood considering only K nearest distances. The value of K is tuned via crossvalidation for stable and accurate predictions.

Also referred to as web interface, this is an artistic method based on the Streamlit framework offers our two primary groups of audience, patients and healthcare givers, a low input entry page for form submission where they can upload their laboratory results in structured format, as CSV or Excel.

REFERENCES

[1] "Cardiovascular diseases (CVDs)," World Health Organization, 2021. [Online]. Available: https://www.who.int/news-room/factsheets/detail/cardiovascular-diseases-(cvds)

[2] "Heart Disease Facts," Centers for Disease Control and Prevention, 2021. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm

[3] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" PLOS ONE, vol. 12, no. 4, p. e0174944, 2017.

[4] S. V. S. Pakhomov, S. J. Jacobsen, C. G. Chute, and V. L. Roger, "Agreement between patient-reported symptoms and their documentation in the medical record," The American Journal of Managed Care, vol. 14, no. 8, pp. 530–539, 2008.

[5] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," Briefings in Bioinformatics, vol. 19, no. 6, pp. 1236–1246, 2018.

[6] R. Smith, "An Overview of the Tesseract OCR Engine," in Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 2, 2007, pp. 629–633.

[7] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.

[8] J. A. Damen et al., "Prediction models for cardiovascular disease risk in the general population: systematic review," BMJ, vol. 353, p. i2416, 2016.

[9] S. Swaminathan et al., "A machine learning approach to triaging patients with chronic obstructive pulmonary disease," PLOS ONE, vol. 12, no. 11, p. e0188532, 2017.

[10] D. R. Gonzalez, T. Carpenter, J. I. van Hemert, and J. Wardlaw, "An´ open source toolkit for medical imaging de-identification," European Radiology, vol. 20, no. 8, pp. 1896–1904, 2010.

[11] M. Shouman, T. Turner, and R. Stocker, "Using data mining techniques in heart disease diagnosis and treatment," in 2012 Japan-Egypt Conference on Electronics, Communications and Computers, 2012, pp. 173– 177.

[12] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," Procedia Technology, vol. 10, pp. 85–94, 2013.

[13] Z. Yang, Y. Zhang, X. Ding, Y. Cong, Z. Xiao, and J. Wu, "A Web-based System for Cardiovascular Disease Risk Assessment Using Machine Learning Techniques," in 2019 IEEE International Conference on Healthcare Informatics (ICHI), 2019, pp. 1–6.

[14] A. M. Alaa, J. Yoon, S. Hu, and M. van der Schaar, "Personalized risk scoring for critical care prognosis using mixtures of Gaussian processes," IEEE Transactions on Biomedical Engineering, vol. 65, no. 1, pp. 207– 218, 2017.