



PERIPHERAL ARTERIAL DISEASE DETECTION USING MACHINE LEARNING

Ayesha Sayed¹, Ayishath Raheesha², Isha M Bhandary³, Ishita P Acharya⁴,

Ms. Shwetha Kamath⁵

Student, Dept. of Computer Science & Engineering, Mangalore Institute of Technology & Engineering, Moodabidri, India^{1,2,3,4}

Assistant Professor, Dept. of Computer Science & Engineering, Mangalore Institute of Technology & Engineering, Moodabidri, India⁵

Abstract: Peripheral Arterial Disease (PAD) is a serious vascular condition that is often underdiagnosed due to reliance on manual interpretation of Doppler ultrasound reports. Early detection is critical to prevent complications such as critical limb ischemia and amputation. This study presents an automated PAD detection system that integrates Optical Character Recognition (OCR) with machine learning to analyze Color Doppler reports. The system extracts key clinical parameters, including the Ankle-Brachial Index (ABI), using text parsing techniques from reports in PDF and image formats. A hybrid decision framework combining a Random Forest classifier with rule-based clinical logic is employed to determine PAD presence and severity. A clinical override rule ensures that PAD is flagged when $ABI \leq 0.90$, improving diagnostic reliability. To address class imbalance, Synthetic Minority Oversampling Technique (SMOTE) is applied, resulting in improved accuracy, precision, recall, and F1-score. The proposed system reduces manual effort, enhances diagnostic consistency, and enables real-time clinical decision support, making it suitable for practical healthcare applications.

Keywords: Peripheral Arterial Disease (PAD), Machine Learning, Optical Character Recognition (OCR), Random Forest, Ankle-Brachial Index (ABI), SMOTE, Healthcare Automation, Doppler Report Analysis.

I. INTRODUCTION

Peripheral Arterial Disease (PAD) is a common yet often underdiagnosed vascular disorder caused by the narrowing of arteries, leading to reduced blood flow to the lower limbs. If left undetected, PAD can progress to severe complications such as critical limb ischemia, non-healing ulcers, and even amputation. Early diagnosis plays a crucial role in improving patient outcomes; however, traditional diagnostic approaches rely heavily on manual interpretation of Doppler ultrasound reports and calculation of the Ankle-Brachial Index (ABI). These methods are time-consuming, prone to human error, and may lead to inconsistent results, especially when handling large volumes of patient data. With advancements in artificial intelligence and machine learning, automated systems have emerged as effective solutions for improving diagnostic accuracy and efficiency in healthcare. Machine learning models are capable of analyzing complex clinical data and identifying patterns that may not be easily detected through manual methods. However, many existing systems depend on structured datasets and do not effectively handle unstructured medical reports such as scanned Doppler images. To address these limitations, this research proposes a web-based automated PAD detection system that integrates Optical Character Recognition (OCR) with machine learning techniques. The system extracts key clinical parameters from Doppler reports and applies a hybrid decision mechanism using a Random Forest classifier combined with rule-based clinical logic. A clinical safety feature ensures that PAD is flagged when ABI values are less than or equal to 0.90. By automating report analysis and providing real-time results, the proposed system aims to reduce manual workload, enhance diagnostic consistency, and support timely clinical decision-making.

The main contributions of this research include the development of an end-to-end automated PAD detection system that integrates Optical Character Recognition (OCR) with machine learning techniques. The system efficiently handles unstructured Doppler reports through advanced text extraction and parsing methods, enabling accurate feature identification. A hybrid decision mechanism is implemented by combining a Random Forest classifier with rule-based clinical logic to enhance prediction reliability. Additionally, an ABI-based override rule is incorporated to ensure clinical safety and prevent misclassification of critical cases. The system is further deployed as a web-based application, enabling real-time PAD detection and severity classification for practical healthcare use.



II. LITERATURE SURVEY

The application of machine learning in detecting Peripheral Arterial Disease (PAD) has been widely explored through various approaches, including clinical data analysis, waveform interpretation, and multimodal systems. Sasikala and Mohanarathinam (2024) developed a machine learning-based PAD detection system using datasets from the UCI Machine

Learning Repository. Their approach involved preprocessing, normalization, and feature selection combined with Hyperband optimization. To address class imbalance, SMOTE was applied. Multiple classifiers such as Enhanced Decision Tree, Support Vector Machine (SVM), Random Forest, K-Nearest Neighbour, XGBoost, and Logistic Regression were evaluated, achieving accuracy above 95%. However, the system required high computational resources and introduced potential noise due to oversampling [1].

Austin et al. (2022) compared Logistic Regression with Random Forest models using clinical datasets. The Random Forest model outperformed traditional methods by capturing nonlinear relationships among clinical variables such as age and comorbidities. The study also demonstrated the ability of Random Forest to evaluate feature importance, though challenges like overfitting and reduced interpretability were observed [2].

Normahani et al. (2022) focused on Doppler waveform analysis for PAD detection. They extracted statistical features such as mean, standard deviation, skewness, and kurtosis, along with time-frequency features using Discrete Wavelet Transform. The SVM classifier achieved approximately 86% accuracy, while Logistic Regression reached about 88%. This study highlighted the effectiveness of feature-engineered approaches but relied on structured waveform data [3].

Al-Ramini et al. (2022) explored PAD detection using laboratory-based gait data. They analyzed ground reaction force and joint kinematics using machine learning models such as Neural Networks, Random Forest, SVM, and Logistic Regression. Random Forest achieved accuracy of up to 89%, demonstrating the potential of non-invasive biomarkers for disease detection. However, the approach requires specialized equipment, limiting practical applicability [4].

Sonnenschein et al. (2021) introduced an AI-based scoring system that uses standard clinical and laboratory parameters to identify PAD severity and the need for urgent vascular intervention. Their Random Forest-based model showed strong predictive performance and a positive correlation with ABI values, emphasizing the importance of integrating clinical indicators in machine learning systems [5].

Flores et al. (2021) reviewed the role of artificial intelligence and machine learning in improving PAD detection and treatment. The study highlighted the effectiveness of algorithms such as Random Forest, SVM, and Gradient Boosting in combining data from multiple sources like imaging, blood flow measurements, and electronic health records. However, challenges such as algorithmic bias, lack of validation, and data heterogeneity were identified [6].

Jana et al. (2020) proposed a Doppler spectrogram-based computer-aided diagnosis system for PAD detection. The system extracted hemodynamic features such as Pulsatility Index and Spectral Broadening Index and used machine learning classifiers. The SVM model achieved high accuracy of approximately 97.91%, demonstrating strong performance in distinguishing normal and diseased arterial conditions [7].

Ross et al. (2016) developed machine learning models using diverse clinical, demographic, and genomic data to predict PAD and mortality risk. The Random Forest model achieved an AUC of 0.87, outperforming traditional logistic regression models. This study demonstrated the capability of machine learning to analyze complex and heterogeneous datasets for improved prediction [8].

From the reviewed literature, it is evident that most existing approaches rely on structured datasets or specialized data acquisition methods. However, limited work has been done on handling unstructured Doppler reports using OCR integrated with machine learning. Additionally, many systems lack clinically interpretable decision mechanisms. These gaps motivate the development of the proposed system, which focuses on real-world applicability by combining OCR, machine learning, and rule-based clinical validation.



III. METHODS

The proposed system for Peripheral Arterial Disease (PAD) detection is designed as a comprehensive pipeline that integrates Optical Character Recognition (OCR), data preprocessing, machine learning, and rule-based clinical decision logic. The system processes unstructured Color Doppler reports and transforms them into structured data for automated diagnosis. The methodology consists of several stages, including data acquisition, preprocessing, feature extraction, model development, decision logic integration, system deployment, and evaluation.

A. Data Collection and Input Handling

The system is designed to accept Color Doppler ultrasound reports as the primary input source. These reports are commonly generated in clinical settings and contain essential diagnostic information such as blood flow measurements and AnkleBrachial Index (ABI) values. The reports may be available in multiple formats, including PDF, JPEG, and PNG, which makes it necessary for the system to handle diverse input types. Users, including clinicians and technicians, upload these reports through a web-based interface. The interface is designed to be simple and user-friendly, ensuring that users with minimal technical expertise can easily interact with the system. Once the report is uploaded, the system performs initial validation to ensure that the file format is supported and that the file is not corrupted. Handling unstructured clinical reports presents significant challenges due to variations in formatting, layout, and quality. Different hospitals and diagnostic centers use different report templates, which can include variations in font styles, table structures, and annotation formats. Therefore, the system is designed to be flexible and robust in handling such variability. The dataset was divided into training, validation, and testing sets in the ratio of 70:20:10. This ensured proper evaluation of the model's generalization capability across unseen data.

B. Preprocessing and Image Enhancement

Preprocessing is a critical step that prepares the input data for accurate text extraction and analysis. Since the uploaded reports may vary in quality, resolution, and orientation, preprocessing ensures consistency and improves the performance of subsequent stages. If the uploaded file is in PDF format, it is first converted into image format using appropriate libraries. The system then applies several image enhancement techniques, including noise reduction, contrast enhancement, and normalization. These steps help in removing unwanted artifacts and improving the visibility of textual content. Another important aspect of preprocessing is ensuring a standard resolution, typically around 300 dpi, which improves OCR accuracy. Additionally, techniques such as grayscale conversion and edge detection may be applied to highlight textual regions within the image. These preprocessing steps reduce variability in the input data and ensure that the OCR module can accurately extract information.

C. Optical Character Recognition (OCR) and Text Extraction

The preprocessed images are then passed to the OCR module, which is responsible for extracting textual information from the reports. The system uses OCR engines such as Tesseract to convert image-based text into machine-readable format. The extracted text may include patient details, clinical measurements, and diagnostic observations. However, OCR output is often unstructured and may contain errors due to image quality issues or variations in formatting. To address this, the system employs text parsing techniques to identify and extract relevant information. Special attention is

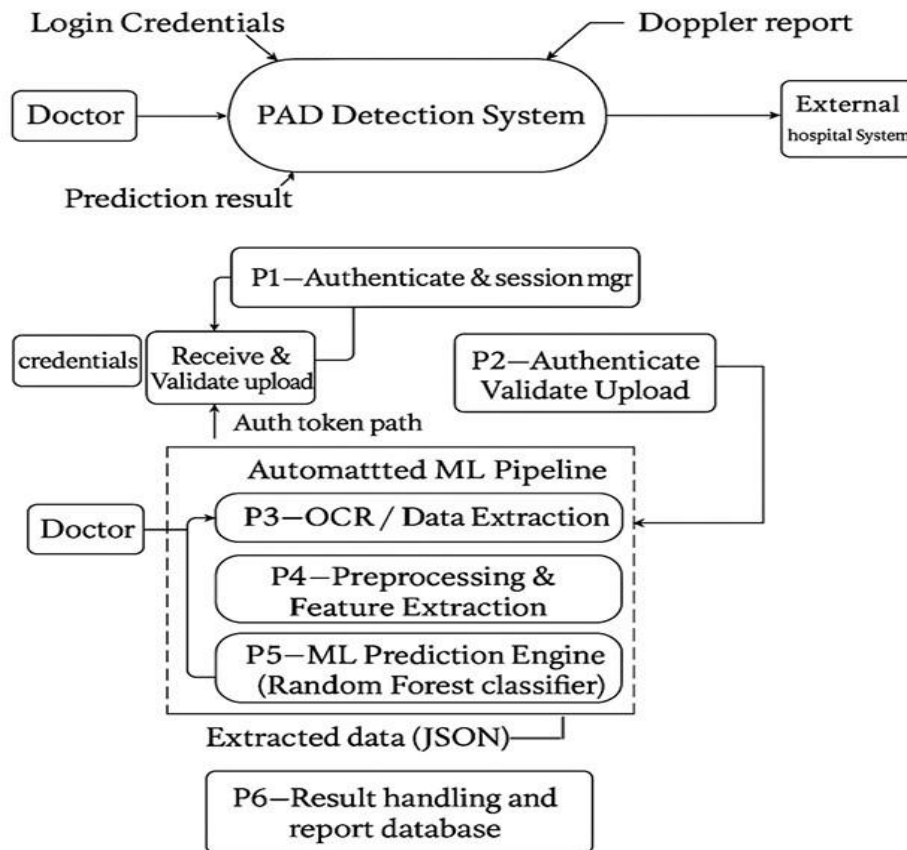


Figure.1: Data Flow Diagram of the Proposed PAD Detection System

given to identifying key clinical parameters such as the Ankle-Brachial Index (ABI), which is a crucial indicator for PAD diagnosis. The system uses pattern-matching techniques to locate ABI values within the extracted text, even when they are presented in different formats. This ensures that important diagnostic information is accurately captured despite variations in report structure. The accuracy of OCR extraction is influenced by factors such as image resolution, formatting variability, and noise. To mitigate errors, preprocessing techniques and pattern-based validation were applied. However, minor extraction inaccuracies may still occur in low-quality reports.

D. Feature Extraction and Data Structuring

Once the relevant information is extracted, the system converts the unstructured text into a structured format suitable for machine learning analysis. This involves organizing the extracted data into predefined fields such as patient details, ABI values, and other clinical parameters. The structured data is stored in formats such as JSON or CSV, which allows for easy manipulation and analysis. Feature extraction focuses on identifying the most relevant parameters that contribute to PAD detection. Among these, the ABI value is considered the most critical feature. The system also accounts for inconsistencies in report formats by using flexible parsing rules. For example, ABI values may be represented in different ways, such as "ABI Right: 0.82" or "Rt ABI = 0.82." The parsing module is designed to handle such variations and ensure accurate extraction of features.

E. Machine Learning Model Development

The core of the proposed system is the machine learning model used for PAD detection and severity classification. A Random Forest classifier is selected due to its robustness, ability to handle nonlinear relationships, and effectiveness in handling structured clinical data. The model is trained using labeled datasets that include different PAD severity levels such as Normal, Mild, Moderate, and Severe. During training, the model learns patterns and relationships between input features and output classes. One of the major challenges in training the model is class imbalance, where certain severity



levels are underrepresented in the dataset. To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) is applied. SMOTE generates synthetic samples for minority classes, resulting in a more balanced dataset. The application of SMOTE significantly improves the model's ability to correctly classify underrepresented categories. As observed in the results, the model's performance improved from an accuracy of 68% to 75%, with notable improvements in precision, recall, and F1score. Although other models such as Support Vector Machine and Logistic Regression were considered as baseline approaches, Random Forest was selected due to its superior performance in handling nonlinear relationships, robustness to noise, and ability to work effectively with structured features extracted from OCR outputs.

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.ensemble import RandomForestClassifier
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE
from imblearn.pipeline import Pipeline as ImbPipeline
import joblib
```

Figure 2: Random Forest Model Training and Architecture Overview

F. Hybrid Decision Logic and Clinical Rule Integration

To ensure reliable and clinically valid predictions, the system incorporates a hybrid decision-making approach. This approach combines machine learning predictions with rule-based clinical logic. The Random Forest model provides predictions regarding the presence and severity of PAD. However, machine learning models may occasionally produce incorrect predictions, especially in cases with low confidence or ambiguous data. To address this limitation, a clinical override rule is implemented. According to this rule, if the extracted ABI value is less than or equal to 0.90, the system automatically classifies the case as PAD, regardless of the model's prediction. This rule is based on established clinical guidelines and ensures that critical cases are not overlooked. The integration of rule-based logic with machine learning enhances the reliability, interpretability, and safety of the system.

G. System Deployment and Web-Based Implementation

The entire system is implemented as a web-based application using the Flask framework. The web application serves as the interface between users and the backend processing modules. The backend handles tasks such as file upload, preprocessing, OCR, feature extraction, and model inference. The system also manages user authentication to ensure secure access. Once the analysis is complete, the results are displayed on a user-friendly dashboard. The output includes extracted parameters, predicted PAD status, severity classification, and confidence indicators. Additionally, the system generates debug logs for each session, which can be used for auditing and improving system performance. The modular architecture of the system allows for easy updates and scalability. New features, improved models, or enhanced OCR techniques can be integrated without significant changes to the overall system.

H. Model Evaluation and Performance Analysis

The performance of the system is evaluated using standard classification metrics such as accuracy, precision, recall, and



F1score. These metrics provide a comprehensive understanding of the model's effectiveness in detecting PAD and classifying its severity. The results show that the application of SMOTE significantly improves model performance, particularly in handling minority classes. The confusion matrix analysis indicates that the model performs well in both binary classification (PAD vs No PAD) and multiclass classification (severity levels). Although some misclassifications occur between adjacent severity classes, the overall performance demonstrates that the system is reliable and suitable for practical deployment in clinical environments. The evaluation results indicate that the model performs reliably in both binary and multiclass classification tasks. The improvement after applying SMOTE highlights the importance of handling class imbalance in medical datasets. The presence of false negatives indicates that certain PAD cases exhibit borderline clinical features, making them difficult to classify. This is a common challenge in medical diagnosis systems where feature overlap exists. However, the integration of the ABI-based rule significantly reduces the risk of missing critical cases, thereby improving clinical safety.

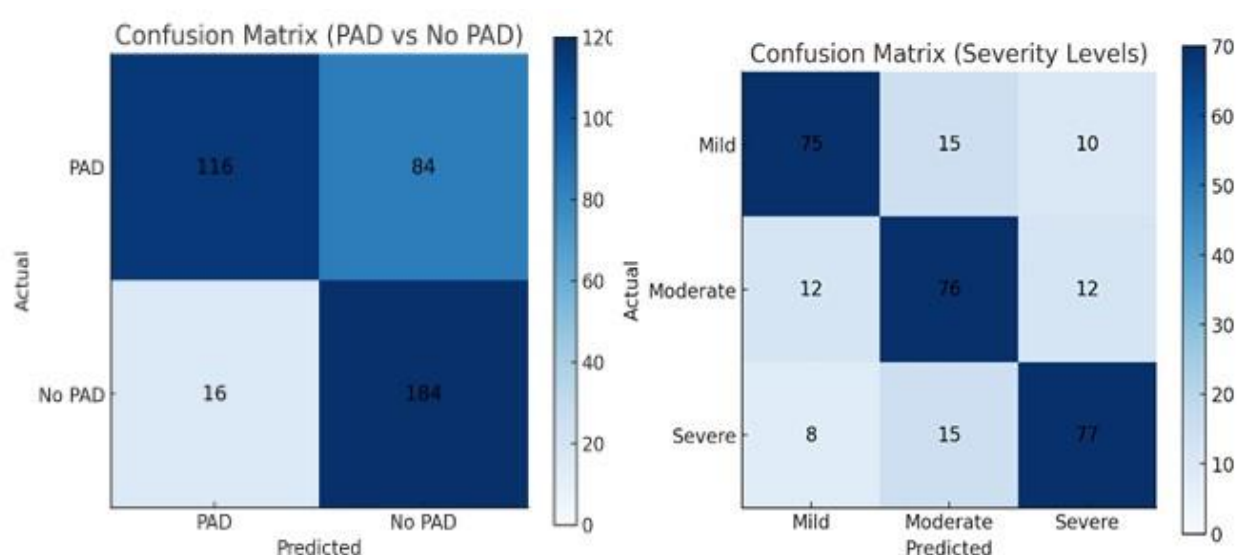


Figure 3: Confusion Matrix Showing PAD Classification Performance

Metric	Before SMOTE	After SMOTE
Accuracy	68%	75%
Precision	75%	87%
Recall	63%	75%
F1-score	58%	66.8%

Table.1: Performance Comparison Before and After Applying SMOTE



IV. SYSTEM ARCHITECTURE

The system architecture of the proposed Peripheral Arterial Disease (PAD) detection system illustrates the interaction between various modules, including the user interface, preprocessing components, machine learning model, and decision logic. The system is designed as a web-based application that enables users to upload Color Doppler reports and receive automated diagnostic results. Initially, the user uploads a Doppler report through the web interface. The uploaded file is passed to the preprocessing module, where operations such as file validation, image conversion, noise reduction, and contrast enhancement are performed. This ensures that the input data is suitable for accurate text extraction. The processed image is then forwarded to the Optical Character Recognition (OCR) module, which extracts textual information from the report. The extracted text is analyzed using a parsing module that identifies key clinical parameters, particularly the Ankle-Brachial Index (ABI). These parameters are structured into a format suitable for machine learning analysis. The structured data is then passed to the Random Forest classifier, which predicts the presence and severity of PAD. To ensure clinical reliability, a rule-based decision module is integrated with the machine learning model. If the extracted ABI value is less than or equal to 0.90, the system automatically flags PAD regardless of the model's prediction. The final output, including PAD status, severity classification, and extracted details, is displayed on the user interface. The overall system architecture ensures efficient data flow, modular design, and seamless integration of OCR and machine learning components, enabling accurate and real-time PAD detection.

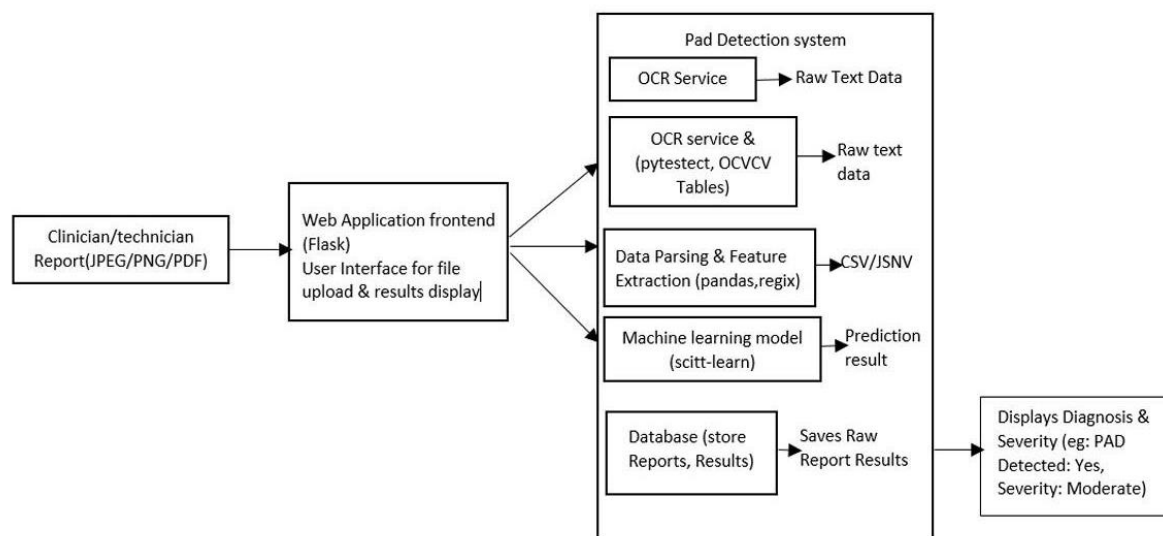


Figure 4: System Architecture of the Proposed PAD Detection System

V. RESULTS AND DISCUSSION

The performance of the proposed Peripheral Arterial Disease (PAD) detection system was evaluated based on its ability to accurately classify PAD presence and severity using the integrated OCR and machine learning pipeline. The system was tested within the web-based environment, where users uploaded Color Doppler reports, and the system performed real-time analysis to generate diagnostic results. One of the major challenges encountered during model development was class imbalance in the dataset. Initially, the dataset contained uneven distribution among PAD severity classes, with fewer samples in the Mild and Moderate categories. This imbalance caused the Random Forest classifier to bias towards majority classes, resulting in reduced performance. Before applying Synthetic Minority Oversampling Technique (SMOTE), the model achieved an accuracy of approximately 68%, with a precision of 75%, recall of 63%, and F1-score of 58%. To address this issue, SMOTE was applied during the training phase to generate synthetic samples for underrepresented classes. This significantly improved the balance of the dataset and enhanced the model's learning capability. After applying SMOTE, the model's performance improved to an accuracy of 75%, precision of 87%, recall of 75%, and F1-score of 66.8%. These results demonstrate a clear improvement in the model's ability to correctly classify minority classes and reduce misclassification errors. The confusion matrix further illustrates the effectiveness of the proposed system. In the binary classification scenario (PAD vs No PAD), the model correctly identified 116 PAD cases



and 184 non-PAD cases, while misclassifying 84 PAD cases as non-PAD and 16 non-PAD cases as PAD. This indicates that the model achieves high precision in detecting PAD, although some false negatives remain, which is a common challenge in medical diagnosis systems. In the multiclass classification scenario, the system demonstrated strong performance in categorizing PAD severity levels. The model correctly classified 75 Mild, 76 Moderate, and 77 Severe cases. Most misclassifications occurred between adjacent severity levels, such as Mild being classified as Moderate or Moderate being classified as Severe. This behavior reflects the natural overlap in clinical characteristics between severity stages, rather than a fundamental limitation of the model. The integration of OCR with machine learning proved to be effective in processing unstructured Doppler reports. The system successfully extracted relevant clinical parameters and provided consistent outputs through the web application. Additionally, the rule-based clinical override ensured that cases with ABI values less than or equal to 0.90 were correctly flagged as PAD, improving the reliability of the system. Overall, the results indicate that the proposed system performs efficiently in both detection and severity classification of PAD. The combination of SMOTE, Random Forest, and rule-based logic enhances the robustness and clinical applicability of the system, making it suitable for real-world healthcare environments.

Overall, the results demonstrate that the combination of OCR, SMOTE, Random Forest, and rule-based logic creates a robust and reliable system capable of handling real-world variability in medical reports.



Figure. 5: Detection result showing Severe Severity

VI. CONCLUSION

The proposed Peripheral Arterial Disease (PAD) detection system demonstrates the effective integration of Optical Character Recognition (OCR) and machine learning techniques for automated diagnosis using Color Doppler reports. By utilizing a Random Forest classifier combined with rule-based clinical logic, the system is capable of accurately detecting PAD and classifying its severity into multiple categories. The inclusion of the Ankle-Brachial Index (ABI) override rule ensures adherence to clinical standards and enhances the reliability of predictions. The application of Synthetic Minority Oversampling Technique (SMOTE) significantly improved the performance of the model, particularly in handling class imbalance and enhancing the detection of minority severity classes. Experimental results indicate notable improvements in accuracy, precision, recall, and F1-score, confirming the robustness of the system. The web-based implementation provides a user-friendly interface for uploading reports and obtaining real-time diagnostic results, reducing manual workload and minimizing human error. The system also ensures consistency in interpretation and supports faster clinical decision-making. Overall, the proposed system offers a scalable, efficient, and reliable solution for PAD detection, with potential applications in clinical settings, especially in environments where rapid and automated diagnosis is essential. Future work will focus on improving model accuracy using advanced machine learning and deep learning techniques such as Gradient Boosting and Neural Networks. Additionally, integrating more diverse datasets and enhancing OCR capabilities using advanced APIs can further improve system performance. The system can also be extended for integration with hospital Electronic Health Record (EHR) systems to support large-scale clinical deployment.



REFERENCES

- [1] P. Sasikala and A. Mohanarathinam, "A powerful Peripheral Arterial Disease detection using machine learning-based severity level classification model and hyperparameter optimization methods," *Biomedical Signal Processing and Control*, vol. 85, pp. 105341, 2024. DOI: 10.1016/j.bspc.2023.105341
- [2] A. M. Austin, N. Ramkumar, B. Gladders, J. A. Barnes, M. A. Eid, K. O. Moore, M. W. Feinberg, M. A. Creager, M. Bonaca, and P. P. Goodney, "Using a cohort study of diabetes and peripheral artery disease to compare logistic regression and machine learning via random forest modeling," *BMC Medical Research Methodology*, vol. 22, no. 1, pp. 1–11, 2022. DOI: 10.1186/s12874-022-01774-8
- [3] P. Normahani, V. Sounderajah, D. Mandic, and U. Jaffer, "Machine learning-based classification of arterial spectral waveforms for the diagnosis of peripheral artery disease in the context of diabetes: A proof-of-concept study," *Computers in Biology and Medicine*, vol. 146, pp. 105695, 2022. DOI: 10.1016/j.combiomed.2022.105695
- [4] A. Al-Ramini, M. Hassan, F. Fallahtafte, M. A. Takallou, H. Rahman, B. Qolomany, I. I. Pipinos, F. Alsaleem, and S. A. Myers, "Machine Learning-Based Peripheral Artery Disease Identification Using Laboratory-Based Gait Data," *Sensors*, vol. 22, no. 19, pp. 7432, 2022. DOI: 10.3390/s22197432
- [5] K. Sonnenschein, S. D. Stojanović, N. Dickel, J. Fiedler, J. Bauersachs, T. Thum, M. Kunz, and J. Tongers, "Artificial Intelligence Identifies an Urgent Need for Peripheral Vascular Intervention by Multiplexing Standard Clinical Parameters," *Journal of Clinical Medicine*, vol. 10, no. 19, pp. 4532, 2021. DOI: 10.3390/jcm10194532
- [6] A. M. Flores, F. Demsas, N. J. Leeper, and E. G. Ross, "Leveraging Machine Learning and Artificial Intelligence to Improve Peripheral Artery Disease Detection, Treatment, and Outcomes," *Circulation Research*, vol. 128, no. 12, pp. 1794–1812, 2021. DOI: 10.1161/CIRCRESAHA.121.319282
- [7] B. Jana, K. Oswal, S. Mitra, and S. Banerjee, "Detection of peripheral arterial disease using Doppler spectrogram based expert system for Point-of-Care applications," *Biomedical Signal Processing and Control*, vol. 55, pp. 101637, 2020. DOI: 10.1016/j.bspc.2019.101637
- [8] E. G. Ross, N. Shah, R. Dalman, K. Nead, J. Cooke, and N. J. Leeper, "The use of machine learning for the identification of peripheral artery disease and future mortality risk," *Journal of Vascular Surgery*, vol. 64, no. 5, pp. 1515–1522.e3, 2016. DOI: 10.1016/j.jvs.2016.05.113