Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Data-Driven Approach for Non-Invasive Hemoglobin Prediction and Anemia Classification

Vijay. P[1], Mehfooza. M[2,*]
[1]*Student, School of Computer Science and Engineering, Vellore Institute of Technology, India.*
[2]*Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology, India.*
**mehfooza.m@vit.ac.in*

ARTICLE INFO

Received: 24 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Introduction: Anemia is a global health concern that affects millions of people worldwide and is primarily diagnosed through invasive blood tests. These methods, while accurate, are costly, time-consuming, and inaccessible in many remote areas. Non-invasive solutions using machine learning and physiological signals such as Photoplethysmography (PPG) offer a promising alternative.

Objective: This study aims to develop a machine learning-based approach for non-invasive hemoglobin estimation and anemia classification. The goal is to evaluate the feasibility of using vital signs (SpO_2, heart rate) and PPG signals to provide an accessible and cost-effective solution for real-time anemia detection.

Methods: The research utilized a dataset comprising patient vitals and hemoglobin levels. Empirical formulas were applied to compute PPG signals, and machine learning models were trained on these features. Two approaches were analyzed: one based on vital signs alone and another incorporating PPG signals. Models such as Gradient Boosting, Random Forest, and Linear Regression were evaluated using metrics like R^2 score, RMSE, and MAE.

Results: The PPG-based approach achieved near-perfect accuracy with an R^2 score of 1.000, while the vitals-based approach showed high practical applicability with an R^2 score of 0.971 using Gradient Boosting. Anemia classification models achieved up to 99.54% accuracy, demonstrating the effectiveness of machine learning for non-invasive diagnostics.

Conclusion: This study highlights the potential of machine learning in non-invasive hemoglobin estimation and anemia detection. While PPG-based models offer superior accuracy, vitals-based models are more practical for real-world applications due to sensor limitations. The findings pave the way for integrating non-invasive anemia detection into wearable devices and healthcare systems.

Keywords: Non-invasive hemoglobin estimation, anemia detection, machine learning, Photoplethysmography (PPG), SpO_2, heart rate, real-time monitoring, healthcare technology.

INTRODUCTION

Anemia is a prevalent global health concern, affecting millions and potentially leading to severe complications if left untreated. It is characterized by a deficiency in hemoglobin, the oxygen-carrying protein in red blood cells. Common symptoms include fatigue, dizziness, shortness of breath, and cognitive impairment, while severe cases can result in organ damage and an increased risk of mortality. Traditional hemoglobin measurement relies on invasive blood tests, which, although accurate, are time-consuming and often inaccessible in remote or underdeveloped regions. These challenges highlight the need for a portable, non-invasive, and real-time approach to anemia detection.

Despite being the gold standard for hemoglobin measurement, traditional blood tests present several limitations. The procedure is invasive and may cause discomfort or infection risks. It requires laboratory facilities, trained professionals, and specialized equipment, making it costly and difficult to access in certain regions. Additionally, the time delay in processing blood samples prevents real-time monitoring, and the need for frequent hemoglobin checks becomes inconvenient due to repeated blood draws.

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Non-invasive solutions utilizing photoplethysmography (PPG) sensors, vital signs, and machine learning offer a promising alternative for anemia screening, providing a more accessible and efficient approach to hemoglobin estimation. Optical sensor-based methods, such as the system developed by Timm et al., enable real-time, pain-free monitoring, reducing infection risks and facilitating immediate clinical interventions [1,14].

This study aims to develop and evaluate a machine learning-based non-invasive hemoglobin estimation system by analyzing the relationship between PPG signals, vital signs, and hemoglobin levels. The proposed approach involves building predictive models for hemoglobin estimation and anemia classification, while also comparing machine learning-based predictions with empirical methods to assess their accuracy and reliability. Furthermore, this research explores the feasibility of integrating such a system into real-time healthcare applications to improve accessibility and efficiency in anemia detection.

This research investigates the feasibility of using PPG signals and vital signs for hemoglobin estimation by developing machine learning models trained on PPG data, SpO_2, and heart rate to predict hemoglobin levels. The accuracy of ML-based predictions is evaluated against empirical methods to determine their effectiveness. Additionally, this study explores the potential for real-time, non-invasive anemia detection and its applications in healthcare settings. The findings demonstrate that a data-driven approach to anemia detection can serve as a cost-effective, accessible, and portable alternative to traditional hemoglobin testing methods, potentially improving early diagnosis and treatment for individuals in remote or resource-limited areas.

OBJECTIVES

The primary objective of this study is to develop and evaluate machine learning models for non-invasive hemoglobin estimation and anemia classification. By leveraging physiological signals such as SpO_2, heart rate, and computed PPG values, the research aims to compare the predictive accuracy of multiple regression and classification techniques. Additionally, the study seeks to identify the most practical approach for real-world implementation, balancing predictive performance with sensor availability and feasibility in non-clinical settings.

LITERATURE REVIEW

Non-invasive hemoglobin estimation has emerged as a promising alternative to traditional blood sampling, aiming to reduce patient discomfort and enable real-time monitoring. In recent years, various machine learning and deep learning techniques have been explored to extract meaningful information from physiological signals such as photoplethysmography (PPG), SpO_2, and heart rate. However, many studies have focused on single algorithms or specific data acquisition methods, leaving a gap in comprehensive comparative analysis across multiple models.

Several studies have utilized deep learning to predict hemoglobin concentration from imaging data. For instance, Khan et al. proposed a real-time non-invasive hemoglobin prediction system using deep learning-enabled smartphone imaging [2,13]. While innovative, this method relied on facial image analysis, making it susceptible to variations in ambient lighting and motion artifacts. Similarly, another study introduced an integrated deep learning solution for estimating heart rate and SpO_2 at laboratory-level accuracy [3,15]. Although effective for vital sign monitoring, this approach did not explore machine learning models for hemoglobin prediction or provide a systematic model comparison.

In contrast, studies using embedded platforms have focused on multi-wavelength PPG signals. Pinto et al. developed a five-wavelength PPG-based system for non-invasive hemoglobin measurement [4,11]. While demonstrating practical feasibility, this work did not evaluate multiple machine learning models using vital sign data such as SpO_2 and heart rate. Another study explored hemoglobin concentration measurement through fingertip PPG signals [5,14], but its approach remained limited to a single predictive model without assessing the benefits of comparative machine learning techniques [16].

A significant study on non-invasive hemoglobin estimation using embedded platforms proposed a method that derived hemoglobin values based on multi-wavelength PPG signals processed through an empirical formula, correlating light absorption with hemoglobin concentration. This approach effectively bypassed the need for invasive blood sampling, offering a promising alternative for real-time anemia detection. However, the study primarily relied on pre-defined equations rather than exploring various machine learning models to optimize accuracy. The authors

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

also highlighted challenges in obtaining stable AC and DC components of PPG signals in real-world conditions, a limitation that restricts the direct applicability of such models without additional sensor advancements.

Additionally, remote vital sign monitoring frameworks, such as ReViSe, have utilized smartphone cameras to capture physiological signals from facial images [6,12]. While ReViSe offers an end-to-end solution for remote health monitoring, it primarily focuses on heart rate, SpO_2, and blood pressure, without directly addressing hemoglobin estimation through comparative machine learning analysis.

By addressing these gaps, this study aims to evaluate multiple machine learning models for hemoglobin estimation, integrating both PPG and vital sign data to enhance prediction accuracy. Furthermore, the study seeks to determine the most practical approach for real-world implementation, balancing predictive performance with sensor availability and feasibility of data acquisition in non-clinical settings.

METHODOLOGY

The data set used in this study consists of 102 patient records obtained from hospital lab reports. The collected features include SpO_2, heart rate, hemoglobin levels, and anemia status, which serve as the primary inputs for our machine learning models. Notably, the dataset primarily comprises individuals in their early 20s, mostly students, leading to a relatively homogeneous age distribution. While this ensures consistency in physiological variations, it may also limit the generalizability of the model to broader age groups.

However, since photoplethysmography (PPG) readings were not available, PPG660 and PPG940 values were computed using an empirical formula. These equations estimate PPG signals based on the molar extinction coefficients of hemoglobin at specific wavelengths. The general form of the equation [1] , as presented in [4] is given by:

$$PPG^{\lambda} = \varepsilon HbO_2^{\lambda} \cdot HbO_2 + \varepsilon Hb^{\lambda} \cdot Hb \quad (1)$$

where $\varepsilon HbO_2^{\lambda}$ is the molar extinction coefficient for oxy-hemoglobin at a given wavelength λ, and εHbλ is the molar extinction coefficient for deoxy-hemoglobin at the same wavelength. For our study, we specifically used the following equations for the 660 nm and 940 nm wavelengths:

$$PPG660 = 319.6 \times HbO2 + 3226.56 \times Hb \quad (2)$$

$$PPG940 = 1214 \times HbO2 + 693 \times Hb \quad (3)$$

The molar extinction coefficient values used for these calculations were obtained from Prahl, S. [7]. Since hemoglobin exists in oxygenated (HbO_2) and deoxygenated (Hb) forms, and as oxygen transport in the blood primarily occurs via hemoglobin [8], their respective values were derived as follows:

$$HbO_2 = Hemoglobin \times (SpO_2/100) \quad (4)$$

$$Hb = Hemoglobin \times (1 - (SpO_2/100)) \quad (5)$$

By integrating these calculations, we were able to reconstruct the missing PPG660 and PPG940 values for the dataset. For hemoglobin estimation, two different approaches were explored. The first approach, a vitals-based method, relies only on SpO_2 and heart rate as input features, eliminating the need for PPG signals and making it compatible with standard pulse oximeters. The second approach incorporates PPG660 and PPG940 values along with SpO_2 and heart rate to enhance prediction accuracy by leveraging the additional physiological information provided by PPG signals.

To evaluate both methods, multiple machine learning models were trained, including Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, XGBoost, LightGBM, Support Vector Regression (SVR), and MLP Regressor. The models were assessed using performance metrics such as R^2 score, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to determine the most effective approach for hemoglobin estimation.

For real-time implementation, a two-wavelength PPG sensor was used to obtain SpO_2 and heart rate readings. Since this sensor does not provide direct PPG signal values, the vitals-based approach was selected for real-time, non-

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

invasive anemia detection. This method ensures a practical alternative to traditional blood tests while maintaining accessibility and ease of use.

RESULTS

The evaluation of the developed models for hemoglobin estimation and anemia classification was conducted using the dataset obtained from lab records. The performance of the models was analyzed using multiple regression and classification techniques. The results of each approach are detailed below.

Hemoglobin Estimation

Two distinct approaches were used for hemoglobin estimation:The first approach utilized SpO_2 and heart rate as input features to predict hemoglobin levels. This method was chosen for its practical applicability, as these parameters can be directly obtained from commercially available sensors.

The second approach estimated hemoglobin levels using computed PPG values, derived using an empirical formula. Since these PPG values were calculated based on hemoglobin itself, this method exhibited higher accuracy, as it inherently preserved the relationship embedded in the equations.The performance of both vitals-based and PPG-based models was assessed using R^2 score, RMSE, and MAE. The results are summarized below:

Table 1- Vitals model accuracy scores

Vitals based models' accuracy score	
Model	R^2 Score
Gradient Boosting	0.971
LightGBM	0.967
Random Forest	0.964
Decision Tree	0.956
XGBoost	0.951
SVR	0.926
MLP Regressor	0.925
Linear Regression	0.915
Ridge Regression	0.915
Lasso Regression	0.913

Table 2- PPG model accuracy scores

PPG Ratio based models' accuracy score	
Model	R^2 Score
Linear Regression	1.000
Ridge Regression	1.000
Lasso Regression	0.9999
Random Forest	0.9997
Decision Tree	0.9996
XGBoost	0.9992
Gradient Boosting	0.9990
SVR	0.9985
LightGBM	0.9983
MLP Regressor	0.7425

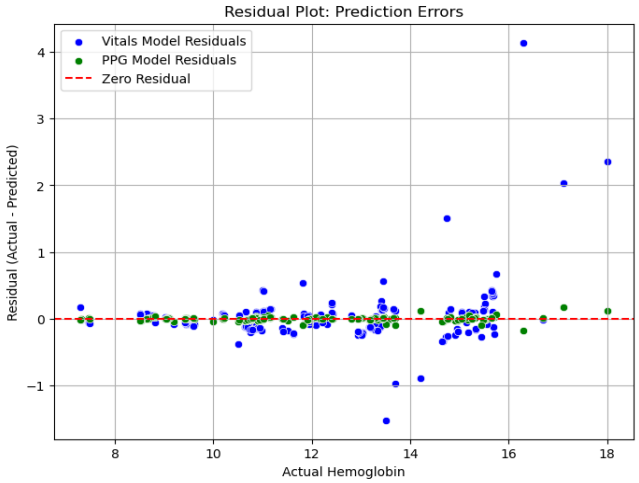

Figure 1– Residual Plot: Prediction Errors of Vitals-based and PPG-based Models

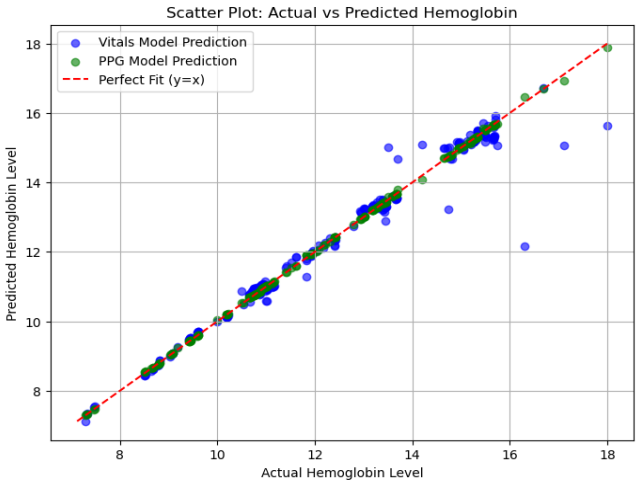

Figure 2 - Residual Plot: Prediction Errors of Vitals-based and PPG-based Models

Journal of Information Systems Engineering and Management
2025, 10(39s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The vitals-based approach achieved its highest performance with the Gradient Boosting model, which obtained an R^2 score of 0.971, RMSE of 0.398, and MAE of 0.126. In contrast, the PPG-based approach recorded near-perfect accuracy with the Linear Regression model, achieving an R^2 score of 1.000, RMSE of 2.684×10^{-15}, and MAE of 2.122×10^{-15}.

Figures below illustrate the prediction performance:

Figure 1: Residual Plot showing Prediction Errors of Vitals-Based and PPG-Based Models. Figure 2: Scatter Plot comparing Actual vs Predicted Hemoglobin Levels.An anemia classification model was developed using hemoglobin levels and gender as input features, based on clinically established thresholds (13.5 g/dL for males and 12 g/dL for females). The results are summarized below:

Table 3- PPG model accuracy scores

Anemia Classification Model Performance	
Model	**Accuracy**
Gradient Boosting	99.54%
LightGBM	99.54%
Random Forest	99.07%
AdaBoost	99.07%
XGBoost	99.07%
Decision Tree	98.61%
KNN	98.61%
SVM	94.91%
Naive Bayes	93.98%
Logistic Regression	91.67%

DISCUSSIONS

The results indicate that PPG-based models consistently outperformed vitals-based models in terms of accuracy, as expected due to their inherent reliance on hemoglobin-dependent equations for computing PPG values.The scatter plot and residual plot demonstrate that PPG-based models exhibit minimal residual errors and a near-perfect fit with actual hemoglobin levels ($R^2 = 1$). However, despite their superior accuracy, practical implementation of PPG-based models is challenging due to limitations in commercial sensors, which do not provide direct AC/DC components of PPG signals for each wavelength.

In contrast, the vitals-based approach (using SpO_2 and heart rate) is more feasible for real-world applications because these parameters can be measured by standard commercial sensors widely available in medical devices, consumer health gadgets, and wearable technology.For anemia classification, the Gradient Boosting model demonstrated exceptional performance with an accuracy of 99.54%, correctly identifying almost all anemic and non-anemic cases based on clinically established thresholds.

While non-invasive hemoglobin estimation offers a painless and real-time alternative to traditional blood sampling, practical implementation depends heavily on advancements in sensor technology to provide direct AC/DC components for multi-wavelength PPG signals.Future developments in sensor technology could further enhance the accuracy and feasibility of non-invasive hemoglobin estimation systems in real-world settings, enabling higher precision in anemia detection through wearable devices or remote health monitoring platforms.

CONCLUSION

This study demonstrates the feasibility of non-invasive hemoglobin estimation and anemia classification using machine learning models trained on physiological signals such as SpO_2, heart rate, and computed PPG values. The results indicate that PPG-based models achieve superior accuracy due to their inherent reliance on hemoglobin-dependent equations, while vitals-based models offer a practical solution for real-world applications due to the availability of commercial sensors.

Research Article

Anemia classification models further validate the effectiveness of machine learning, with the Gradient Boosting classifier achieving an accuracy of 99.54%, showcasing its reliability in diagnosing anemia based on clinically established thresholds. These findings highlight the potential of integrating non-invasive hemoglobin monitoring into wearable devices and consumer health technologies, enabling real-time anemia detection and improving accessibility to healthcare.Future advancements in sensor technology could enhance the precision of non-invasive hemoglobin estimation by providing direct AC/DC PPG signal components, bridging the gap between theoretical accuracy and practical implementation. This research paves the way for innovative solutions in remote health monitoring, offering a painless and cost-effective alternative to traditional blood tests.

REFERENCES

[1] Challenges of traditional hemoglobin measurement and the role of non-invasive techniques. (n.d.). IEEE Sensors Journal. Retrieved from [https://doi.org/10.1109/JSEN.2023.3284895]
[2] Real-time non-invasive hemoglobin prediction using deep learning-enabled smartphone imaging. (n.d.). National Center for Biotechnology Information. Retrieved from [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11218390/]
[3] Introduction of Integrated Image Deep Learning Solution and how it brought laboratorial level heart rate and blood oxygen results to everyone. (n.d.). arXiv Preprint. Retrieved from [https://arxiv.org/abs/2204.07999]
[4] Non-invasive hemoglobin measurement using embedded platform. (2020). Sensing and Bio-Sensing Research, 30, 100370. Retrieved from [https://doi.org/10.1016/j.sbsr.2020.100370]
[5] Non-Invasive Measurement of Heart Rate and Hemoglobin Concentration Level Through Fingertip. (n.d.). ResearchGate. Retrieved from [https://www.researchgate.net/publication/281372564_Non-Invasive_Measurement_of_Heart_Rate_and_Hemoglobin_Concentration_Level_Through_Fingertip]
[6] ReViSe: Remote Vital Signs Measurement Using Smartphone Camera. (n.d.). arXiv Preprint. Retrieved from [https://arxiv.org/abs/2206.08748]
[7] [Physiology, Oxygen Transport - StatPearls - NCBI Bookshelf. (n.d.). National Center for Biotechnology Information. Retrieved from [https://www.ncbi.nlm.nih.gov/sites/books/NBK538336/]
[8] [Optical properties of hemoglobin and blood. (n.d.). Optical Microscopy and Laser Cytometry Group. Retrieved from [https://omlc.org/spectra/hemoglobin/summary.html]
[9] World Health Organization (2011). Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity. WHO Guidelines on Nutrition. Retrieved from [https://apps.who.int/iris/handle/10665/85839]
[10] Signal processing techniques for wearable photoplethysmography sensors. (n.d.). IEEE Sensors Conference. Retrieved from [https://doi.org/10.1109/ICSENS.2009.5398321]
[11] M. Mehfooza & V. Pattabiraman (2018): SP-DDPT: a simple prescriptive based domain data preprocessing technique to support multilabel-multicriteria learning with expert information, International Journal of Computers and Applications, DOI: 10.1080/1206212X.2018.1547475.
[12] Nageswara Prasadhu, Mehfooza.M (2020),"An Efficient Hybrid Load Balancing Algorithm for Heterogeneous Data Centers in Cloud Computing",International Journal of Advanced Trends in Computer Science and Engineering,Vol.9,No.3,pp:3078-3085.
[13] Mehfooza, M. and Pattabiraman, V. (2021) 'A new efficient learning approach E-PDLA in assessing the knowledge of breast cancer dataset', Int. J. Services and Operations Management, Vol. 38, No. 2, pp.153–160.
[14] Mehfooza M, I HB., (2021). An automated prescriptive domain data preprocessing algorithm to support multilabel-multicriteria classification for Indian coastal dataset, crop dataset, and breast cancer dataset. Int J Commun Syst. 2021;e4796. https://doi.org/10.1002/dac.4796.
[15] P. Kuppusamy, M. M. Basha and C. -L. Hung, "Retinal Blood Vessel Segmentation using Random Forest with Gabor and Canny Edge Features," 2022 International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN), Villupuram, India, 2022, pp. 1-4, doi: 10.1109/ICSTSN53084.2022.9761339.
[16] Garapati, R., Munavar Basha, M. (2023). A Systematic Review on Recommender System Models, Challenges, Domains and Its Perspectives. In: Nandan Mohanty, S., Garcia Diaz, V., Satish Kumar, G.A.E. (eds) Intelligent Systems and Machine Learning. ICISML 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 470. Springer, Cham. https://doi.org/10.1007/978-3-031-35078-8_38.