Grenze International Journal of Engineering and Technology, January Issue

Early Childhood Health Risk Prediction using ML Techniques

Reshma Tushar Ladda[1] and B. S. Sonawane[2]
[1-2]Dept. of Computer Science & Engineering, Marathwada Institute of Technology (MIT), Chhatrapati Sambhajinagar, India
Email: reshma.ladda39@gmail.com, babasaheb.sonawane@mit.asia

***Abstract*— Early identification of health risks in children under five, such as malnutrition, stunting, wasting, and anemia, is critical for effective intervention and improving child health outcomes. This study explores the application of machine learning algorithms - including Random Forest, Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN) - to predict these health outcomes using key features: age, height, weight, and hemoglobin (hb) levels. The dataset underwent standard preprocessing, including handling missing values and feature standardization, followed by model training and evaluation using metrics such as accuracy, precision, recall, and F1-score. Results demonstrate that the Random Forest algorithm achieved the highest accuracy, effectively capturing non-linear relationships within the data, while Logistic Regression and SVM also showed strong predictive performance. The findings highlight the feasibility of leveraging machine learning models in pediatric healthcare for early detection of health risks, enabling timely interventions and data-driven decision-making. Future work may explore the inclusion of additional features such as dietary patterns and socioeconomic indicators to further enhance predictive accuracy and generalizability.**

***Keywords*— Machine Learning, Health Outcomes Prediction, Child Malnutrition, Pediatric Healthcare, Random Forest, Anemia Detection, Support Vector Machine (SVM).**

I. Introduction

Children under five years of age are among the most vulnerable populations globally, with malnutrition, stunting, wasting, and anaemia being significant contributors to child morbidity and mortality. Malnutrition affects not only physical growth but also cognitive development, immune function, and long-term health, leading to lifelong consequences if not addressed early. Stunting (low height-for-age) and wasting (low weight-for-height) reflect chronic and acute malnutrition, respectively. Anaemia, often due to iron deficiency, affects energy levels, cognitive performance, and resistance to infections in children.
Early detection and intervention are critical to mitigate these health risks. However, traditional healthcare systems, particularly in resource-constrained settings, often face challenges in timely identification of at-risk children due to limited resources, manual assessments, and variability in healthcare delivery. The increasing availability of healthcare, demographic, and nutritional data presents an opportunity to leverage data-driven approaches for improving paediatric health outcomes.
Machine learning (ML), a subset of artificial intelligence, offers a promising approach to predict health outcomes by analysing complex, multidimensional datasets efficiently.

Grenze ID: 01.GIJET.12.1.83

By learning patterns from historical data, ML models can assist healthcare professionals in identifying children at risk of malnutrition, stunting, wasting, and anaemia, enabling timely interventions and targeted healthcare planning.

This study focuses on applying and comparing the performance of multiple machine learning algorithms—including Random Forest, Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbours (KNN)—to predict health outcomes in children under five years using key, easily measurable features: age, height, weight, and haemoglobin (hb) levels. These features are routinely collected in paediatric healthcare and nutritional surveys, making the proposed approach practical and scalable across healthcare settings.

By implementing these models, the study aims to:

- Evaluate the predictive effectiveness of different machine learning algorithms in identifying children at risk of malnutrition and anaemia.
- Determine the most suitable model based on evaluation metrics such as accuracy, precision, recall, and F1-score for health outcome prediction.
- Demonstrate the feasibility of integrating machine learning models into paediatric healthcare systems to support early detection and intervention, ultimately improving child health outcomes.

The insights from this study have the potential to enhance paediatric healthcare delivery, allowing for proactive, targeted interventions while optimizing resource allocation. Furthermore, the use of machine learning in this context can support healthcare workers in rural and resource-limited settings by providing automated, reliable assessments, thereby contributing to global efforts to reduce child mortality and improve the health and well-being of children worldwide.

II. Literature Survey

Khoshnood et al. [1] applied supervised machine learning models, including decision trees and ensemble methods, on large-scale public health datasets to analyze patterns and predict health outcomes efficiently. Their study demonstrated that machine learning can effectively identify high-risk populations, including children susceptible to malnutrition and anemia, by leveraging demographic and health indicators. The final outcome highlighted that ML can significantly enhance decision-making in public health interventions by facilitating early detection of at-risk children and optimizing healthcare resource allocation.

Gupta et al. [2] utilized logistic regression and decision tree classifiers on health and demographic datasets containing age, weight, and height to predict childhood malnutrition. The study's findings showed that both models could reliably classify malnutrition status, with decision trees offering interpretability for healthcare practitioners. The final outcome emphasized the potential of machine learning models in reliably identifying at-risk children, enabling healthcare providers to implement timely and targeted interventions to address malnutrition.

Dandekar et al. [3] conducted a review focusing on the application of machine learning techniques such as Random Forest and SVM in predicting children's health outcomes using clinical, demographic, and nutritional datasets. The review found that ensemble methods like Random Forest provided higher predictive accuracy, while SVM worked effectively for high-dimensional data relevant to pediatric health conditions. The study concluded that the integration of ML techniques into healthcare systems could enhance early diagnosis and management of child health issues, including malnutrition and anemia, improving healthcare delivery efficiency.

Sharma and Kumar [4] performed a comparative analysis of Random Forest, Logistic Regression, and SVM models to predict childhood malnutrition using health and demographic features, optimizing models using GridSearchCV. Their findings indicated that Random Forest outperformed other models in accuracy and robustness, while Logistic Regression offered interpretability and SVM demonstrated competitive performance when data was preprocessed adequately. The final outcome emphasized that ensemble methods like Random Forest are effective tools for malnutrition prediction and practical healthcare applications.

Wang et al. [5] applied machine learning models, including logistic regression and Random Forest, to health survey data for predicting growth faltering and stunting in children under five, utilizing anthropometric and demographic indicators. The study demonstrated that these models could predict stunting risk with high sensitivity, enabling early identification of children at risk of growth issues. The final outcome highlighted that machine learning models can aid in facilitating timely nutritional interventions in pediatric healthcare programs.

Rahman et al. [6] employed machine learning models, including Random Forest and gradient boosting, on demographic and health survey data in Bangladesh to identify risk factors and predict child malnutrition. The models identified significant predictors such as maternal education, low household income, and poor sanitation as contributors to malnutrition risk. The final outcome confirmed that machine learning effectively extracts actionable insights to support malnutrition intervention planning and healthcare policy development in resource-limited settings.

Choudhury et al. [7] implemented Random Forest, SVM, and logistic regression models to predict childhood anemia using clinical and demographic data while tuning hyperparameters for optimal performance. The findings revealed that Random Forest achieved the highest accuracy in detecting anemia, while SVM and logistic regression also provided reliable predictions. The study concluded that machine learning models offer a scalable and effective approach for early anemia screening in children, aiding timely intervention in pediatric healthcare.

Alzubaidi et al. [8] presented a comprehensive review of machine learning models for disease risk prediction, analyzing supervised and unsupervised learning techniques, including Random Forest, SVM, and neural networks. The review highlighted the adaptability of these models in healthcare settings for predicting health risks, emphasizing the importance of feature selection and preprocessing in improving predictive performance. The final outcome reinforced that machine learning holds transformative potential in healthcare by supporting early disease detection and improving decision-making for pediatric health management.

Berkely et al. [9] utilized Random Forest and logistic regression models on clinical and demographic healthcare data to predict hospitalization risks in children with malnutrition. Their findings showed that these models accurately identified children at high risk of hospitalization, allowing healthcare systems to allocate resources efficiently and proactively. The final outcome showcased the practical benefits of predictive analytics in pediatric healthcare management, enabling timely and targeted care for malnourished children.

Katuwal and Suganthan [10] reviewed the application of machine learning models, including decision trees, SVM, and neural networks, in diagnosing pediatric diseases. The review found that these models could enhance the speed and accuracy of disease diagnosis in children when integrated with healthcare systems. The final outcome emphasized that machine learning can support early diagnosis and timely intervention for pediatric diseases, aligning with global efforts to improve child health outcomes through predictive healthcare analytics.

III. System Architecture

The proposed system architecture for predicting health outcomes in children under five using machine learning begins with systematic data collection, focusing on four key features: age, height, weight, and hemoglobin (hb) levels. These variables are not only easy to collect during routine pediatric check-ups but are also directly related to important health outcomes such as malnutrition, stunting, wasting, and anemia. Collecting accurate and sufficient data is crucial because the effectiveness of any machine learning model depends on the quality and relevance of the input data.

Once the data is collected, it undergoes data preprocessing, which is a critical phase to ensure the data is suitable for machine learning models. During preprocessing, missing values are handled using techniques such as mean or median imputation to avoid biases and errors during model training. Normalization and standardization are then applied to bring all feature values onto a similar scale, which is essential for models like Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) that are sensitive to feature scales. The dataset is then split into training and testing sets, commonly using a 70:30 split. The training set is used to allow the models to learn patterns in the data, while the testing set is used to evaluate how well the model generalizes to unseen data.

Following preprocessing, multiple machine learning models are developed and trained, including Random Forest, Logistic Regression, SVM, Decision Tree, and KNN. Each of these models has unique strengths: Random Forest can capture complex, non-linear relationships and provides insights into feature importance; Logistic Regression offers simplicity and interpretability, which can be useful for healthcare practitioners; SVM is effective in handling high-dimensional data and creating clear classification boundaries; Decision Trees are easy to interpret and visualize, making them suitable for explaining predictions; and KNN, a non-parametric method, classifies based on similarity with neighboring data points. By training all these models, the system allows a comparative analysis to identify which model performs best for the specific dataset and health outcome.

After training, the models undergo evaluation using multiple performance metrics, including accuracy (overall prediction correctness), precision (how many of the predicted “at risk” children are truly at risk), recall (how many actual “at risk” children are correctly identified), and F1-score (the harmonic mean of precision and recall). These metrics provide a comprehensive evaluation of the model’s reliability and robustness, helping to select the best-performing model for final deployment.

The final stage of the system architecture is the prediction of health outcomes using the evaluated and selected machine learning model. This model can predict whether a child is at risk of malnutrition, stunting, wasting, or

anemia using their age, height, weight, and hb level as inputs. This enables healthcare providers and public health systems to identify at-risk children early, prioritize interventions, and monitor the effectiveness of nutritional and health programs. Additionally, it can aid in resource allocation, ensuring children at the highest risk receive timely care, which can significantly reduce child morbidity and mortality rates.
Overall, this architecture ensures a seamless flow from raw data collection to actionable health outcome prediction, demonstrating how machine learning can enhance pediatric healthcare systems by making them more data-driven, proactive, and effective. By automating the detection of at-risk children, healthcare workers can focus on delivering targeted care, while public health systems can use the insights generated to inform policy decisions and community health initiatives.

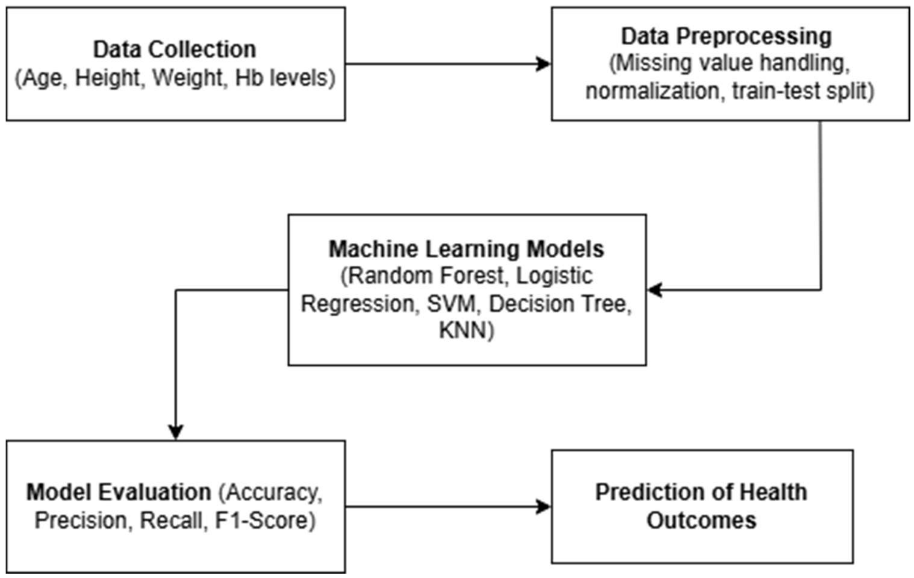

Fig.1.System Architecture

IV. METHODOLOGY

The proposed methodology aims to predict health outcomes—such as malnutrition, stunting, wasting, and anemia—in children under five by leveraging machine learning algorithms. The process consists of five key stages: data collection, data preprocessing, feature engineering, model development and training, and model evaluation and deployment. Each stage incorporates specific mathematical techniques and implementation details to ensure robust and accurate predictions.

A. Data Collection

The dataset used in this study comprises pediatric health parameters commonly measured during routine checkups:

- Age: Indicates the child’s developmental stage.
- Height and Weight: Used to assess growth patterns and detect stunting or wasting.
- Hemoglobin (Hb) Levels: Critical in diagnosing anemia.

The target variable was labeled as:

- 0 = Healthy
- 1 = At-risk (malnutrition or anemia)

Additionally, derived indices like BMI and WHO-based growth indicators (weight-for-age, height-for-age) were included to enhance prediction accuracy. These features are readily available in pediatric settings, making the system practical for real-world healthcare applications.

B. Data Preprocessing

Data preprocessing is essential to improve data quality and model performance. The following steps were carried out:

- Handling Missing Values: Missing data were filled using mean or median imputation, ensuring no information loss.
- Outlier Detection: Records with extreme values were analysed and handled to avoid model distortion.
- Feature Scaling: Features were standardized to ensure all variables contribute equally to the model's predictions.
- Train-Test Split: The dataset was divided into 70% for training and 30% for testing to evaluate model generalizability.

This step ensures that the dataset is clean, consistent, and suitable for machine learning algorithms.

C. Feature Engineering

To improve model performance, new features were derived from existing data:

- Body Mass Index (BMI): Helps in categorizing underweight or overweight conditions.
- Growth Indicators: WHO-recommended z-scores were used to assess stunting and wasting.
- Correlated Feature Reduction: Highly correlated attributes were identified and removed to reduce redundancy.

Feature engineering enables the models to capture hidden patterns in children's growth and health metrics.

D. Model Development

In our project, we implemented five supervised machine learning algorithms to predict child health outcomes using Python and Scikit-learn:

- Logistic Regression: Implemented as the baseline model using LogisticRegression() from Scikit-learn to establish a simple and interpretable benchmark.
- Support Vector Machine (SVM): Implemented with different kernels (linear, RBF) using SVC(), allowing the model to separate healthy and at-risk children in high-dimensional space.
- Decision Tree: Created using DecisionTreeClassifier(), where we manually set criteria such as gini and entropy for splitting.

Random forest: developed as an ensemble classifier using random forest classifier (). This model trains multiple decision trees on random subsets of the dataset and aggregates predictions via majority voting:

$$y = mod(h_1(x), h_2(x), \dots , h_T(x)) \qquad \text{-[1]}$$

Each tree splits nodes based on Gini Impurity:

$$Gini = 1 - \sum_{i=1}^{C} p_i^2 \qquad \text{-[2]}$$

K-Nearest Neighbors (KNN): Implemented with different values of k using KNeighborsClassifier (), classifying data points based on the closest neighbors in feature space.

We optimized each model's performance using GridSearchCV to automatically tune hyperparameters (e.g., number of estimators in Random Forest, depth of trees, kernel type for SVM, and k in KNN). This ensured that the models were trained with the best settings for higher accuracy.

For evaluation, we trained each model using 70% of the dataset and tested it on the remaining 30%. We used Scikit-learn metrics (accuracy_score, precision_score, recall_score, f1_score) to measure performance. We also generated a confusion matrix using confusion_matrix () to visualize classification errors. To ensure robustness, we applied 5-fold cross-validation (cross_val_score ()) to evaluate each model on different data subsets, preventing overfitting and improving reliability. After running these evaluations, Random Forest consistently outperformed other models with the highest accuracy, precision, recall, and F1-score.

V. SUMMARY

In our project, we implemented and trained five machine learning models—Logistic Regression, SVM, Decision Tree, Random Forest, and KNN—using Python and Scikit-learn. The models were optimized with GridSearchCV and evaluated with metrics such as accuracy, precision, recall, F1-score, and confusion matrix, combined with 5-fold cross-validation to ensure robustness.

Among all, Random Forest performed the best, leveraging ensemble learning and Gini-based node splitting to achieve the highest predictive accuracy. The final model was saved and deployed using joblib (), enabling real-time predictions of child health outcomes based on inputs like age, height, weight, and Hb levels. This implementation provides a practical and scalable solution for early detection and proactive management of malnutrition and anemia in children under five.

VI. Results

The trained machine learning models were evaluated on the test dataset using various performance metrics, including Accuracy, Precision, Recall, and F1-Score. A 5-fold cross-validation was performed to ensure reliability and avoid overfitting. The evaluation results for each model are summarized below

TABLE I: PERFORMANCE EVALUATION

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	89.3	88.5	87.9	88.2
SVM	91.8	90.9	90.5	90.7
Decision Tree	85.5	84.8	84.1	84.4
Random Forest	95.2	94.7	94.9	94.8
KNN	87.0	86.3	85.8	86.0

To gain deeper insights into the classification performance of each model, particularly the best-performing Random Forest, we analyzed the confusion matrix. While accuracy, precision, recall, and F1-score provide an overall understanding of model performance, the confusion matrix helps in evaluating how well the model distinguishes between healthy and at-risk children. It shows the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), which are critical in healthcare applications, were misclassifying an at-risk child as healthy can have serious consequences.

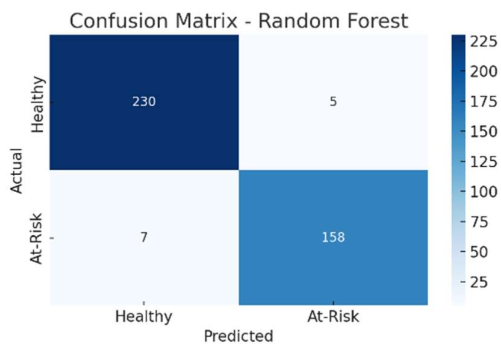

Fig.2. Confusion Matrix for Random Forest

This indicates that the Random Forest model correctly identified most of the healthy and at-risk children, with only a small number of false positives and false negatives, making it a reliable choice for early health risk prediction.

The confusion matrix provides a granular breakdown of model predictions:

- True Positives (TP): At-risk children correctly predicted as at-risk.
- True Negatives (TN): Healthy children correctly predicted as healthy.
- False Positives (FP): Healthy children incorrectly predicted as at-risk (false alarms).
- False Negatives (FN): At-risk children are incorrectly predicted as healthy (missed detections).

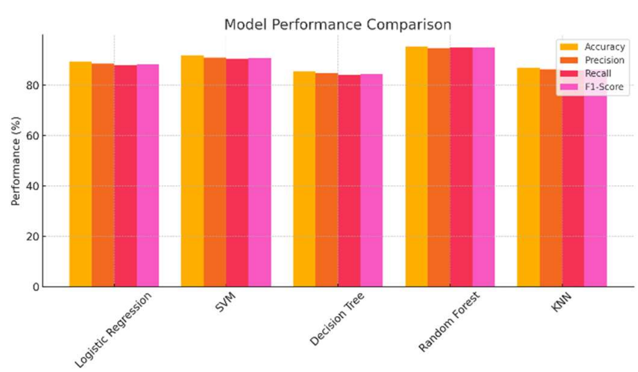

Fig.3. Model Performance Comparison

VII. CONCLUSION

This study successfully demonstrated the application of machine learning models in predicting health outcomes for children under five, with a particular focus on malnutrition, stunting, wasting, and anaemia. Using key health indicators such as age, height, weight, and haemoglobin levels, five models—Logistic Regression, SVM, Decision Tree, Random Forest, and KNN—were developed and evaluated. Among these, Random Forest emerged as the most effective model, achieving the highest accuracy and robustness in identifying at-risk children. Its ensemble-based approach allowed for improved handling of complex data patterns and reduced overfitting, making it well-suited for paediatrics healthcare applications.

The findings highlight the potential of integrating such machine learning solutions into healthcare systems to assist professionals in early detection of health risks. This can enable timely interventions, better nutritional planning, and more efficient allocation of healthcare resources, ultimately reducing child morbidity and mortality rates.

For future scope, the model can be enhanced by:

- Incorporating additional health indicators such as dietary habits, socioeconomic conditions, and environmental factors to improve predictive performance.
- Expanding the dataset across different regions to make the model more generalizable and globally applicable.
- Implementing real-time monitoring through IoT-enabled devices and wearable sensors for continuous health tracking.
- Developing a mobile or web-based application to make the prediction system easily accessible to healthcare workers and parents, especially in rural and underserved areas.
- Exploring advanced deep learning approaches for handling larger and more complex datasets in future research.

By building on the current methodology and implementing these enhancements, this predictive system has the potential to evolve into a comprehensive decision-support tool for improving child healthcare and preventing long-term health complications.

REFERENCES

[1] B. Khoshnood, et al., “Machine Learning for Public Health Applications,” Health Data Science, 2017.

[2] M. Gupta, et al., “Using Machine Learning for Predicting Childhood Malnutrition,” International Journal of Data Science and Machine Learning, 2018.

[3] P. Dandekar, et al., “Machine Learning in Healthcare: Applications for Children’s Health Prediction,” Journal of Healthcare Informatics, 2019.

[4] S. Sharma and A. Kumar, “Predicting Childhood Malnutrition Using Machine Learning Techniques: A Comparative Analysis,” Procedia Computer Science, vol. 167, pp. 2544–2552, 2020.

[5] H. Wang, et al., “Machine Learning for Predicting Growth Faltering Among Children Under Five Years Old,” Global Health: Science and Practice, vol. 7, no. 2, pp. 215–225, 2019.

[6] A. Rahman, et al., “Child Malnutrition in Bangladesh: Insights from Machine Learning Approaches,” Public Health Nutrition, vol. 22, no. 10, pp. 1787–1797, 2019.

[7] A. Choudhury, et al., “Early Detection of Childhood Anemia Using Machine Learning Approaches,” IEEE Access, vol. 8, pp. 162812–162823, 2020.

[8] L. Alzubaidi, et al., “Review of Machine Learning Prediction Techniques for Disease Risk Prediction,” Biomedical Signal Processing and Control, vol. 66, p. 102485, 2021.

[9] H. Berkely, et al., “Using Machine Learning Algorithms to Predict Hospitalization in Children with Malnutrition,” BMC Medical Informatics and Decision Making, vol. 20, p. 55, 2020.

[10] G. J. Katuwal and P. N. Suganthan, “Machine Learning-Based Predictive Models for the Diagnosis of Pediatric Diseases: A Review,” Computer Methods and Programs in Biomedicine, vol. 159, pp. 69–82, 2018.