Health Monitoring System based on LightGBM, CatBoost and Explainable Boost Classifier

Dr. Prithvi .P
Department of Electronics and Communication Engineering
National Institute of Technology Warangal
Telangana, India.
prithvi@nitw.ac.in

Yashvvinie Santhakumar
Department of Electronics and Communication Engineering
National Institute of Technology Warangal
Telangana, India.
yashvvinie2000@gmail.com

Aishwarya Yenmanagandla
Department of Electronics and Communication Engineering
National Institute of Technology Warangal
Telangana, India.
yenmanagandlaaishwarya@gmail.com

Amruthavarshini Sriram
Department of Electronics and Communication Engineering
National Institute of Technology Warangal
Telangana, India.
ruxstudent@gmail.com

***Abstract*—This research introduces a sophisticated health monitoring system that utilizes the vital parameters of Pulse Rate, Temperature, Blood Oxygen Saturation(SPO2), and Respiratory Rate. The system employs advanced machine learning algorithms, including CatBoost, Light Gradient Boosting Machine (LightGBM), and Explainable Boosting Machine (EBM) Classifiers, with the ensemble algorithm designated as Explainable Cat Gradient Boosting Machine (X-CatGBM), for the analysis and interpretation of the collected physiological data. By employing these algorithms, the system achieves real-time and accurate assessments of an individual's health status, providing insights into overall well-being. Notably, the chosen algorithms offer a balance between predictive accuracy and interpretability, ensuring that healthcare professionals and end-users can comprehend the reasoning behind the system's health evaluations. The transparent and interpretable nature of the model outputs enhances trust and understanding, crucial elements for the successful adoption of such technology in healthcare. The study establishes a foundation for the integration of cutting-edge machine learning techniques in health monitoring, holding promise for transformative impacts on remote patient monitoring and personalized healthcare practices.**

***Index Terms*—Pulse-rate, Temperature, SPO2, Respiratory-rate, CatBoost, LightGBM, EBM, X-CatGBM**

I. Introduction

The integration of advanced technologies into healthcare has spurred a transformative shift toward personalized and data-driven approaches. In response to this paradigm, our study introduces a cutting-edge health monitoring system designed to harness the wealth of information embedded in key physiological parameters: Pulse Rate, Temperature, SPO2 (Blood Oxygen Saturation), and Respiratory Rate. These parameters collectively provide a comprehensive snapshot of an individual's vital signs, facilitating a nuanced understanding of their health status. Our aim is to advance the field of health monitoring by deploying state-of-the-art machine learning algorithms—CatBoost, LightGBM, and Explainable Boosting Classifier—thus enabling accurate, efficient, and interpretable health assessments. The resulting ensemble algorithm is referred to as the Explainable Cat Gradient Boosting Machine (X-CatGBM).

The choice of physiological parameters emphasizes a holistic approach to health monitoring. Pulse Rate and Respiratory Rate serve as fundamental indicators of cardiovascular and respiratory health, while Temperature and SPO2 provide insights into metabolic activity and blood oxygenation levels. This multi-parameter approach enhances the system's ability to capture the dynamic interplay of various physiological processes, laying the foundation for a more precise and personalized health monitoring experience.

In tandem with the parameter selection, the adoption of machine learning algorithms addresses the need for robust data analysis. CatBoost, known for its prowess in handling categorical variables, is well-suited for the diverse and complex nature of health data. LightGBM, with its efficiency in processing large datasets and gradient boosting capabilities, complements the system's objective of real-time monitoring. Furthermore, the inclusion of Explainable Boosting Classifier underscores our commitment to transparency and interpretability, ensuring that healthcare professionals and end-users can trust and comprehend the decisions made by the system.

As we delve into the details of our methodology and results, this research seeks to showcase the potential of this integrated health monitoring system. By combining advanced physiological insights with cutting-edge machine learning, we aspire to contribute to a new era of healthcare where proactive and personalized monitoring becomes not only feasible but integral to improving patient outcomes and well-being.

II. Related Works

Pavleen Kaur et al conducted an in-depth examination of prediction models leveraging random forest classifiers and incorporating the IoT across diverse medical conditions, including heart disease, breast cancer, diabetes, thyroid disorders, spect-heart issues, surgical outcomes, dermatological conditions, liver illnesses, and surgical data. The experimental outcomes revealed that the Random Forest machine learning algorithm demonstrated remarkable efficacy, achieving a maximum accuracy of 97.26% when applied to the Dermatology dataset. Notably, IoT devices played a pivotal role in data acquisition and transmission to other systems; however, the investigation addressed the data security concerns within IoT-based frameworks. [2]

Ambrish G et al conducted a study focusing on the utilization of patient data to predict the presence or absence of cardiovascular disease. Employing a UCI dataset, logistic regression (LR) techniques were employed for the categorization of heart illness. Performance enhancement was achieved through pre-processing, feature selection, and training/testing procedures. The LR model demonstrated notable efficacy, achieving an accuracy rate of 87.10% with a training/testing split of 90% and 10%, respectively. [3]

Hari Priya P B et al developed a Health Monitoring System, implemented as a Web Application utilizing the Flask framework and Decision Tree Classification. The system aims to predict patient health levels and identify potential risk areas. The system uses a 93% accuracy score in the prediction process. The proposed method aims to prevent diseases early, enabling better decision-making, trend identification, and efficiency in research and clinical trials. This system also improves healthcare delivery, reduces costs, and accurately handles patient data, ultimately reducing death rates and improving overall health outcomes. [1]

B.Venkataramanaiah et al endeavored to employ biomedical sensors, ARM processors, and FPGA for the detection, testing, analysis, and display of normal or abnormal cases, with the objective of offering cost-effective medical services to individuals residing in remote villages. Through the utilization of the K-Nearest Neighbour (KNN) classifier, their approach surpasses other machine learning algorithms, achieving a maximum accuracy of 99%. [4]

Mehdi Hosseinzadeh et al. had introduced an IoT-based health monitoring system designed for senior citizens who reside independently without familial support. The systematic collection of vital information is facilitated through IoT monitoring devices, and subsequent analysis employs the Multi-Layer Perceptron (MLP) machine learning algorithm. The Decision Tree (J48) model emerged as the most effective technique for predicting health function status, achieving an impressive precision and accuracy rate of 99%. Significantly, the J48 method exhibited the shortest execution time at 12 milliseconds, rendering it well-suited for real-time monitoring applications. [5]

III. Methodology

A. Data Collection, Preprocessing and Problem Formulation:

The dataset used in this research was obtained from Kaggle. It includes vital sign measurements such as pulse rate, body temperature, respiratory rate, and blood oxygen levels. Collection adhered to ethical guidelines, ensuring participant consent and data privacy.

Raw data underwent extensive preprocessing. This process involved handling missing values, outlier detection and removal, and standardization of units for vital sign measurements. Categorical variables were encoded appropriately. Feature selection ensured the inclusion of relevant information.

The study focuses on a binary classification task aimed at identifying emergency states based on vital sign measurements. Emergency conditions are defined when vital signs surpass pre-established thresholds, ensuring the relevance and sensitivity of the classification criteria.

B. Ensemble Model Selection:

1) ***LightGBM****:* LightGBM, known for its efficiency in handling large datasets, was selected as one of the base models. Hyperparameter tuning involved a grid search to optimize parameters, including learning rate, maximum depth, and the number of boosting rounds.

LightGBM constitutes a gradient boosting framework employing tree-based learning algorithms. The mathematical representation of LightGBM encompasses the utilization of a loss function, commonly serving as the objective function undergoing optimization throughout the training process. The objective function can be expressed as shown in Equation 1. The precise structure of the equation is contingent upon the nature of the task at hand, whether regression or classification, and the selected loss function.

$$\text{Objective} = \sum_{i=1}^{n} w_i \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right] + \sum_{i=1}^{n} \frac{\lambda}{2} \sum_{k=1}^{K} w_k^2 \tag{1}$$

where,

- n represents the total number of data points.
- w_i denotes the weight assigned to the i-th data point.
- y_i signifies the true label of the i-th data point, with values of 0 or 1 in the case of binary classification.
- p_i represents the predicted probability that the i-th data point belongs to the positive class.

- λ is the regularization parameter.
- K is the number of leaves in each tree.

This formulation encapsulates the fundamental principles underlying the optimization process for binary classification conducted by LightGBM. It leverages a binary logloss function and incorporates regularization techniques with the objective of enhancing the robustness of the model.

*2) **CatBoost**:* CatBoost, chosen for its robust handling of categorical features, underwent a similar optimization process. Parameters such as tree depth, learning rate, and iterations were fine-tuned for optimal performance.

The mathematical representation of CatBoost involves the optimization of an objective function denoted as Equation 1 throughout the training process. The specific structure of the objective function is contingent upon the nature of the task, be it regression or classification, and the selected loss function. In the context of classification tasks, the logloss (cross-entropy) loss function is commonly employed. The representation of the CatBoost objective function for binary classification is given in the Equation 2.

$$\text{Objective} = \sum_{i=1}^{n} L(y_i, F(x_i)) + \sum_{i=1}^{n} \frac{1}{2} \cdot H(F(x_i)) \quad (2)$$

where

- n: The total number of data points in the dataset.
- $F(x_i)$: The model's prediction for the i-th data point. In the context of gradient boosting, this is the sum of predictions from individual trees.
- y_i: The true label or target value for the i-th data point in the dataset. In binary classification, y_i takes on the values of 0 or 1, indicating the class membership of the corresponding instance.
- $L(y_i, F(x_i))$: The loss function, measuring the difference between the true label y_i and the predicted output $F(x_i)$ for the i-th data point. In binary classification, a common choice is the logloss (cross-entropy) loss function.
- $H(F(x_i))$: The regularization term, which helps prevent overfitting by penalizing complex models. The specific form of this term depends on the regularization strategy used in CatBoost.

The specific formulation of the loss function and regularization term is contingent upon the particular variant of the CatBoost algorithm in use, as well as any customization choices exercised by the user. The overarching objective during the training phase is the minimization of this objective function, with the aim of attaining an effective model.

*3) **Explainable Boost Classifier**:* To enhance interpretability, an Explainable Boost Classifier(EBM) was incorporated into the ensemble. Hyperparameters were carefully chosen to strike a balance between model complexity and interpretability.

InterpretML seamlessly incorporates the EBM, an innovative interpretability algorithm characterized by high explicability and comprehensibility, while maintaining accuracy levels comparable to sophisticated machine learning techniques such as Random Forest and Boosted Trees.

The EBM methodology involves the learning of each feature function, denoted as f_i , utilizing contemporary techniques such as bagging and gradient boosting. It employs a low learning rate to mitigate co-linearity, conducting training iteratively in a round-robin fashion, focusing on one feature at a time. This method elucidates how each feature contributes to the model's prediction, facilitating a comprehensive understanding of the optimal feature function for each feature. The mathematical representation of the EBM classifier is articulated by Equation 3.

$$g(E|y|) = \beta_0 + \sum f_i(x_i) + \sum f_{i,j}(x_i, x_j) \quad (3)$$

where,

- $g(E \mid y|)$: The expected absolute value of the copula density given the observed values.
- β_0: Intercept term in the model.
- $\sum \text{f}_i(x_i)$: Sunmmation over univariate marginal distribution functions $f_i(x_i)$. Each term represents the contribution of a marginal distribution function for a variable x_i in the copula.
- $\sum \text{f}_i, \text{j}(x_i,x_j)$: Summation over bivariate copula functions f_i, j(x_i,x_j). Each term represents the contribution of a bivariate copula function between variables x_i and x_j in the copula.

C. Ensemble Construction:

An ensemble model was formed by combining the predictions of LightGBM, CatBoost, and the Explainable Boost Classifier, designated as Explainable Cat Gradient Boosting Machine (X-CatGBM). A soft voting mechanism was employed, and weights were assigned to each model based on their performance during cross-validation. This approach aimed to maximise predictive accuracy by leveraging the strengths of each model. The combination of the aforementioned classifiers can be accomplished as elucidated below.

METHODOLOGY OF COMBINING ALGORITHMS:

The StackingClassifier is a sophisticated meta-learning algorithm designed to enhance predictive performance and generalizability by amalgamating the outputs of diverse base learners. In the provided code snippet, two gradient boosting algorithms, LightGBM and CatBoost, serve as the base learners. These base learners contribute predictions, which are then utilized by a meta-learner, the Explainable Boosting

Classifier, to construct a meta-model that yields the final prediction for each instance.

LightGBM, known for its leaf-wise tree construction, accelerates training as it prioritizes nodes with the highest information gain. At each leaf node, a value is assigned based on the optimization objective, and these leaf values collectively contribute to the final prediction for a given instance. LightGBM forms an ensemble of such trees, and predictions are derived by summing up the contributions from all trees.

CatBoost, employs an ordered boosting technique that sorts categorical features and utilizes the resulting order for efficient splitting. CatBoost builds symmetric trees, enhancing prediction accuracy and eliminating the need for a separate procedure to find optimal thresholds for each feature. To prevent overfitting and ensure model generalizability, CatBoost incorporates regularization techniques. Predictions are made by aggregating the contributions of all trees within the ensemble.

In contrast, the Explainable Boosting Classifier (EBM) is designed to be interpretable, providing insights into its predictions through a series of decision rules. The additive nature of EBM models means that predictions result from the sum of contributions from different features. Each feature contributes positively or negatively based on its impact, allowing for a clear understanding of the factors influencing the prediction. EBM employs boosting, combining weak learners to form a strong learner. The learning rate controls the contribution of each weak learner, ensuring a balanced and effective ensemble. EBM further facilitates interpretability by allowing the assessment of feature importance, providing insights into which features have a higher impact on predictions.

In X-CatGBM, the integration of LightGBM and CatBoost with EBM involves the generation of predictions by each base estimator based on their respective processes. These predictions are then fed into the Explainable Boosting Classifier, which learns a meta-model. The meta-model determines the optimal weight to assign to each base learner's prediction. The final prediction is a skillful combination of predictions from LightGBM and CatBoost, with weights determined by the learned meta-model from EBM. This collaborative approach not only leverages the strengths of each base learner but also enhances interpretability and transparency in the final predictive model. The StackingClassifier is created with a list of tuples, where each tuple contains the name of the base estimator and the instance of the base estimator. In this case, the list contains two tuples: ('lgbm', lgbm) and ('catboost', catboost). The final_estimator parameter is set to the instance of the meta-learner, which is the Explainable Boosting Classifier.

The stack_method parameter is set to "auto", which means that the StackingClassifier will automatically determine the best way to combine the predictions of the base estimators to make the final prediction. Figure 1 shows how LightGBM, CatBoost and EBM are combined to create the proposed ensembled algorithm, X-CatGBM Algorithm.

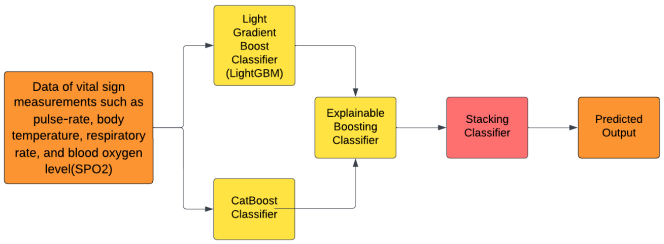

Fig. 1. Methodology of Combining Algorithms

The StackingClassifier will consider various factors such as the performance of each base estimator, the correlation between the base estimators, and the complexity of the problem to determine the best stacking method.

Some of the common stacking methods that the StackingClassifier might use when stack method is set to "auto" include:

- stacking.StackingClassifier.stacking_method.Auto: This method uses a combination of the mean and the standard deviation of the predictions of the base estimators to make the final prediction.
- stacking.StackingClassifier.stacking_method: This method uses the median of the predictions of the base estimators to make the final prediction.
- stacking.StackingClassifier.stacking_method.Average: This method uses the average of the predictions of the base estimators to make the final prediction.
- stacking.StackingClassifier.stacking_method.Majority: This method uses the majority vote of the predictions of the base estimators to make the final prediction.

By setting stack_method to "auto", the StackingClassifier can adapt to the specific problem and choose the best stacking method to improve the accuracy and robustness of the model. For example, if the base estimators have a high correlation, the StackingClassifier might choose the stacking.StackingClassifier.stacking_method.Auto method, which combines the mean and standard deviation of the predictions to make the final prediction. On the other hand, if the base estimators have a low correlation, the StackingClassifier might choose the stacking.StackingClassifier.stacking_method.Median method, which uses the median of the predictions to make the final prediction. In essence, the StackingClassifier excels in aggregating the diverse predictions of its base learners through

the meta-learning process facilitated by the Explainable Boosting Classifier. This adaptability, coupled with the ability to select an optimal stacking method, ensures that the StackingClassifier is well-suited for a wide range of predictive tasks, contributing to improved accuracy and robustness in various machine learning scenarios.

D. Training and Validation:

An 80-20 split was employed to partition the dataset into training and validation sets. The ensemble's generalization performance underwent evaluation through cross-validation with *k=5*. The models underwent training utilizing the designated training set, while a distinct validation set was employed for performance verification. The iterative nature of this process ensured robustness in model training, contributing to the overall reliability of the ensemble.

E. Evaluation Metrics:

An alternative test set is utilized to evaluate the performance of the ensemble model, employing a diverse range of metrics, including:

- Accuracy: Proportion of correctly classified emergency states.
- Precision: Ratio of true positives to all positive predictions.
- Recall: Proportion of actual emergencies correctly identified.
- F1 Score: Harmonic mean of precision and recall.

IV. Results and Discussions

X-CatGBM has demonstrated an impressive accuracy of 99.82% in the health monitor system. This achievement indicates the robustness and effectiveness of the ensemble approach in leveraging the strengths of individual classifiers to enhance overall predictive performance.

The classification reports and confusion matrices of the proposed ensemble algorithm are depicted in Figure 2 and Figure 3.

```
Accuracy: 0.9982349480290253
Confusion Matrix:
[[0.99872351 0.00127649]
 [0.00338409 0.99661591]]
Classification Report:
              precision    recall  f1-score   support

    Abnormal       1.00      1.00      1.00      3917
      Normal       1.00      1.00      1.00      1182

    accuracy                           1.00      5099
   macro avg       1.00      1.00      1.00      5099
weighted avg       1.00      1.00      1.00      5099
```

Fig. 2. Classification Report of the proposed ensemble algorithm

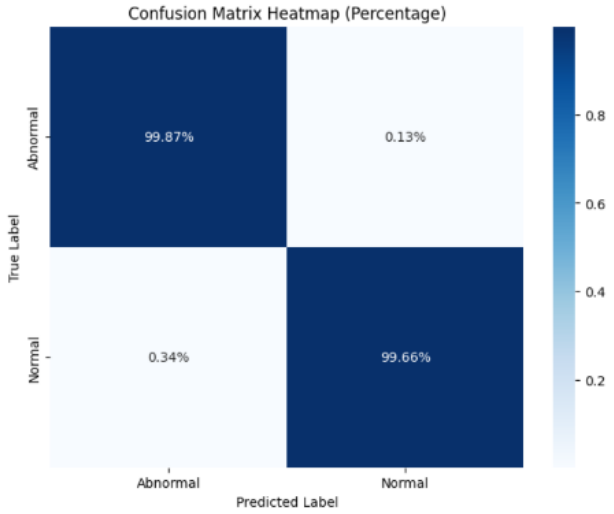

Fig. 3. Confusion Matrix of the proposed ensemble algorithm

In the analysis of the confusion matrix presented above, it is evident that the model has successfully predicted 99.87% of the labels as abnormal (positive class) with accuracy (True Positives), while a mere 0.13% of the labels have been inaccurately classified as abnormal (False Positives).

Furthermore, the model has demonstrated accurate predictions of 99.66% for labels identified as normal (True Negatives), and a minimal 0.34% have been misclassified as abnormal (False Negatives).

These findings contribute to a comprehensive evaluation of the model's classification performance, highlighting its effectiveness in discerning between abnormal and normal instances of an individual's vital signs.

The analysis of accuracy plotted against iterations is illustrated in Figure 4. The graph shows a clear upward trend in accuracy during the early iterations, indicating that the algorithm is effectively learning from the training data. The accuracy curve begins to level off around iteration 40, suggesting that the algorithm's learning rate has slowed down. This plateau could be due to several factors, such as reaching the limits of the model's complexity or exhausting the information available in the training data. Despite the plateau, the algorithm achieves a high level of accuracy, reaching a test accuracy of approximately 0.998. This suggests that the model is capable of making accurate predictions on unseen data.

The shape of the curve suggests that the initial learning rate may have been appropriate, as it led to rapid improvement in accuracy. However, adjustments to the learning rate or other hyperparameters might be considered to potentially overcome the plateau and further improve performance. The minimum

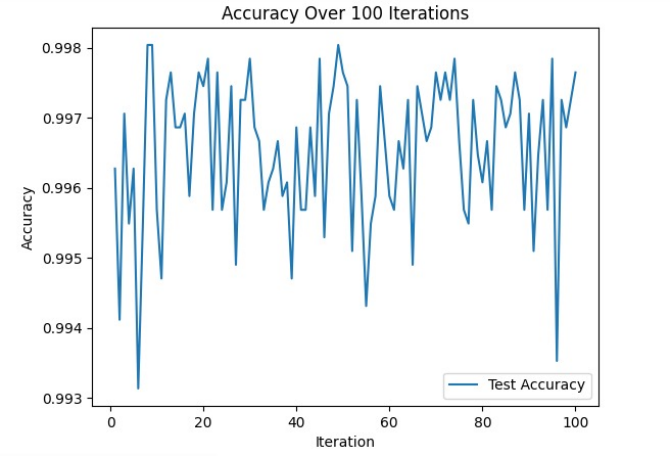

Fig. 4. Analysis of Accuracy against Iterations for X-CatGBM Classifier

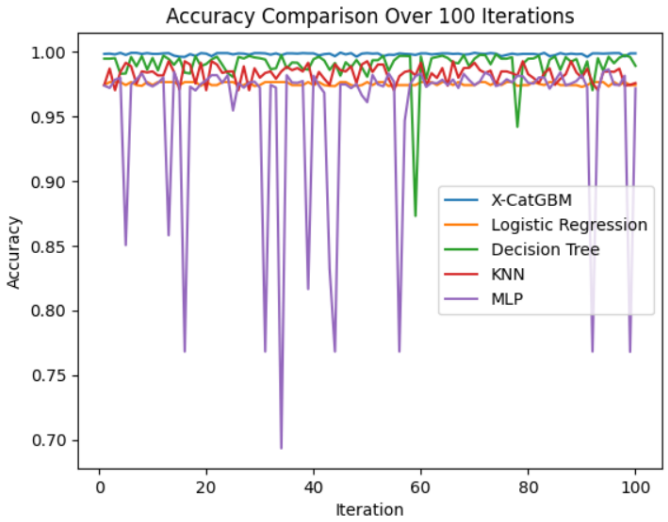

Fig. 5. Graphical Analysis of Accuracy against Iterations for Existing Algorithms and X-CatGBM Classifier

number of iterations required to construct an accuracy curve is contingent upon various factors, such as the model's complexity, dataset size, and the desired convergence criteria. The choice of 100 iterations seems reasonable based on the characteristics of the dataset, the ensemble of algorithms utilized, and the complexity of the model. Further iterations might not yield significant improvements, and could even lead to overfitting or increased computation time.

The accuracy of the model is subject to variation based on the dataset employed. The codes corresponding to the existing binary algorithms, as outlined in the related works, were acquired and subsequently evaluated using the dataset available.

The graphical representation depicted in Figure 5 illustrates the comparative performance in terms of accuracy of five distinct machine learning models across 100 training iterations. Notably, the X-CatGBM model demonstrates a consistent and notably superior performance across all 100 iterations. Its accuracy consistently exceeds 0.993 throughout the training process, potentially nearing a maximal accuracy of nearly 1.00. Conversely, the Logistic Regression, Decision Tree, K-Nearest Neighbours(KNN), and Multi layer Perceptron(MLP) models manifest lower accuracy levels throughout the training iterations. Although their accuracy trajectories exhibit some variability, none of the models are able to exceed the performance exhibited by the X-CatGBM model. It is imperative to acknowledge that the precise accuracy values for the Logistic Regression, Decision Tree, KNN, and MLP models remain somewhat ambiguous from the graph. Collectively, these observations suggest that the X-CatGBM model achieves significantly superior accuracy in comparison to the other evaluated machine learning models. This implies that the X-CatGBM model might represent a more efficacious choice for the specific task or dataset under scrutiny.

The comparative analysis performed on the accuracies of existing binary classification models and X-CatGBM classifier are depicted in the Table 1.

S. No.	Algorithm used	Accuracy Obtained(%)
1	Random Forest Classifier [2]	99.45438839614262
2	Logistic Regression [3]	96.91663494480396
3	Decision Tree Classifier [1]	99.35287400076133
4	K-Nearest Neighbours Classifier [4]	99.61156668105309
5	Multi layer Perceptron Classifier [5]	99.03565537368354
6	X-CatGBM Classifier	99.82349480290253

TABLE I
COMPARISON TABLE BETWEEN EXISTING ALGORITHMS AND X-CATGBM CLASSIFIER

From this table it can be inferred that the X-CatGBM Classifier, outperforms the existing models in binary classification by achieving the highest accuracy which is 99.8%.

V. CONCLUSIONS

In conclusion, our research presents a pioneering health monitoring system integrating Pulse Rate, Temperature, SPO2, and Respiratory Rate, empowered by X-CatGBM. The selected machine learning algorithms ensure efficient data processing with high accuracy and transparent decision-making, fostering trust among healthcare professionals and end-users. This system represents a significant advancement in precision healthcare, offering a comprehensive and interpretable platform for real-time health assessments. While promising, ongoing efforts are needed to address scalability and implementation challenges, paving the way for the integration of advanced technologies in personalized healthcare and ushering in a new era of proactive and tailored medical interventions.

REFERENCES

[1] Hari Priya P B, Dhivya Shree K S, Indhumathi S, Saumya S, and Thangaraj K. Health monitoring using machine learning. International Journal of Progressive Research in Science and Engineering, 2(4):9–12, Apr. 2021.

[2] Kaur.P., Kumar.R. Kumar.M. A healthcare monitoring system using random forest and internet of things (IoT). Multimed Tools Appl 78, 19905–19916 (2019). https://doi.org/10.1007/s11042-019-7327-8.

[3] Ambrish G, Bharathi Ganesh, Anitha Ganesh, Chetana Srinivas, Dhanraj, and Kiran Mensinkal. Logistic regression technique for prediction of cardiovascular disease. Global Transitions Proceedings, 3(1):127–130, 2022. ISSN 2666-285X. doi: https://doi.org/10.1016/j.gltp.2022.04.008. International Conference on Intelligent Engineering Approach(ICIEA-2022).

[4] B Venkataramanaiah and J Kamala. Ecg signal processing and knn classifier-based abnormality detection by vh-doctor for remote cardiac healthcare monitoring. Soft Computing, 24(22):17457–17466, 2020.

[5] Mehdi Hosseinzadeh, Jalil Koohpayehzadeh, Marwan Yassin Ghafour, Aram Mahmood Ahmed, Parvaneh Asghari, Alireza Souri, Hamid Pourasghari, and Aziz Rezapour. An elderly health monitoring system based on biological and behavioral indicators in internet of things. Journal of Ambient Intelligence and Humanized Computing, pages 1–11, 2020.

[6] Celestine Iwendi, Ali Kashif Bashir, Atharva Peshkar, R Sujatha, Jyotir Moy Chatterjee, Swetha Pasupuleti, Rishita Mishra, Sofia Pillai, and Ohyun Jo. Covid-19 patient health prediction using boosted random forest algorithm. Frontiers in public health, 8:357, 2020.

[7] Teuku Rizky Noviandy, Sarah Ika Nainggolan, Raihan Raihan, Isra Firmansyah, and Rinaldi Idroes. Maternal health risk detection using light gradient boosting machine approach. Infolitika Journal of Data Science, 1(2):48–55, 2023.