

# A Two-Stage Anomaly Detection Algorithm for Personalized Healthcare Using Support Vector Machines and Regression Models

Shreea Bose, Venkata Saketh Dakuri, Chittaranjan Hota, Senior Member, IEEE  
Department of Computer Science, BITS Pilani, Hyderabad Campus, Hyderabad, India  
{p20240026, f20220056, hota}@hyderabad.bits-pilani.ac.in

**Abstract**—This paper presents an effective two-stage anomaly detection technique for wireless body area networks (WBANs)-based personalized healthcare monitoring. While standard anomaly detection techniques frequently fail to distinguish between sensor malfunctions and actual medical emergencies, WBANs offer continuous physiological data crucial for detecting irregularities in critical health indicators. Our method uses Support Vector Regression (SVR) for more in-depth contextual analysis and Support Vector Machines (SVM) for initial anomaly categorization. Domain-specific thresholds and sophisticated feature engineering are utilized to accurately identify point anomalies brought on by sensor noise or malfunctions and contextual anomalies resulting from interdependent physiological feature deviations. The robustness of the model is improved by features including heart rate discrepancies, rolling mean, and rolling standard deviation. With 99.61% and 99.97% detection accuracy for point and contextual anomalies, respectively, the model attains an overall accuracy of 99.59%. These findings highlight the promise of scalable, customized anomaly detection systems for WBAN-driven, real-time healthcare, which could lead to prompt treatments and better patient outcomes.

**Index Terms**—SVM, SMO regression, WBANS, rolling mean, Anomaly detection

## I. INTRODUCTION

Wireless Body Area Networks (WBANs) have gained significant traction in healthcare and fitness monitoring, where real-time physiological data collection is crucial. These networks consist of wearable or implantable sensors that continuously track vital signs like heart rate, body temperature, blood oxygen levels, and electrocardiogram (ECG) readings. The primary goal of WBANs is to provide uninterrupted monitoring of an individual's health, enabling early detection of medical conditions, personalized healthcare, and emergency interventions. Out of the advantages of using WBANs, the freedom of movement and noninvasive technology stand out the most. However, since it is a sensor, its functionality faces various challenges, which include security issues, data loss, and network performance issues. The wireless design of WBANs makes data transmission prone to communication disruptions, information loss, or interference from other devices or sensors.

The loss of information or data corruption can lead to severe repercussions, given that health data is in question.

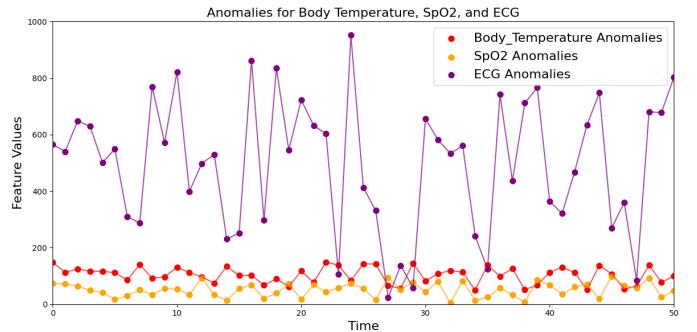


Fig. 1: Plot of Point Anomalies of WBAN features

However, ensuring the accuracy and reliability of these networks is challenging due to the presence of noise, sensor faults, or sudden changes in physiological readings, all of which can generate anomalies in the collected data. Anomaly detection in WBANs is essential to distinguish between normal physiological variations and critical health issues. Identifying these anomalies can help detect irregular heart rhythms, early signs of diseases, or even sensor malfunctions that could otherwise lead to incorrect medical diagnoses. A point anomaly occurs when a single data point deviates from the expected range of values, irrespective of the surrounding context. An example could be, a heart rate reading of 220 bpm in an adult under normal resting conditions. These are context independent and are determined solely by the data point values compared to expected thresholds. A contextual anomaly occurs when a data point deviates from normal patterns given the specific context of the data in a WBAN. The same value may or may not be anomalous depending on contextual factors such as time, activity, or environmental conditions. For example, a heart rate of 150 bpm is normal during exercise while abnormal during sleep.

This work presents a unique two-stage anomaly detection

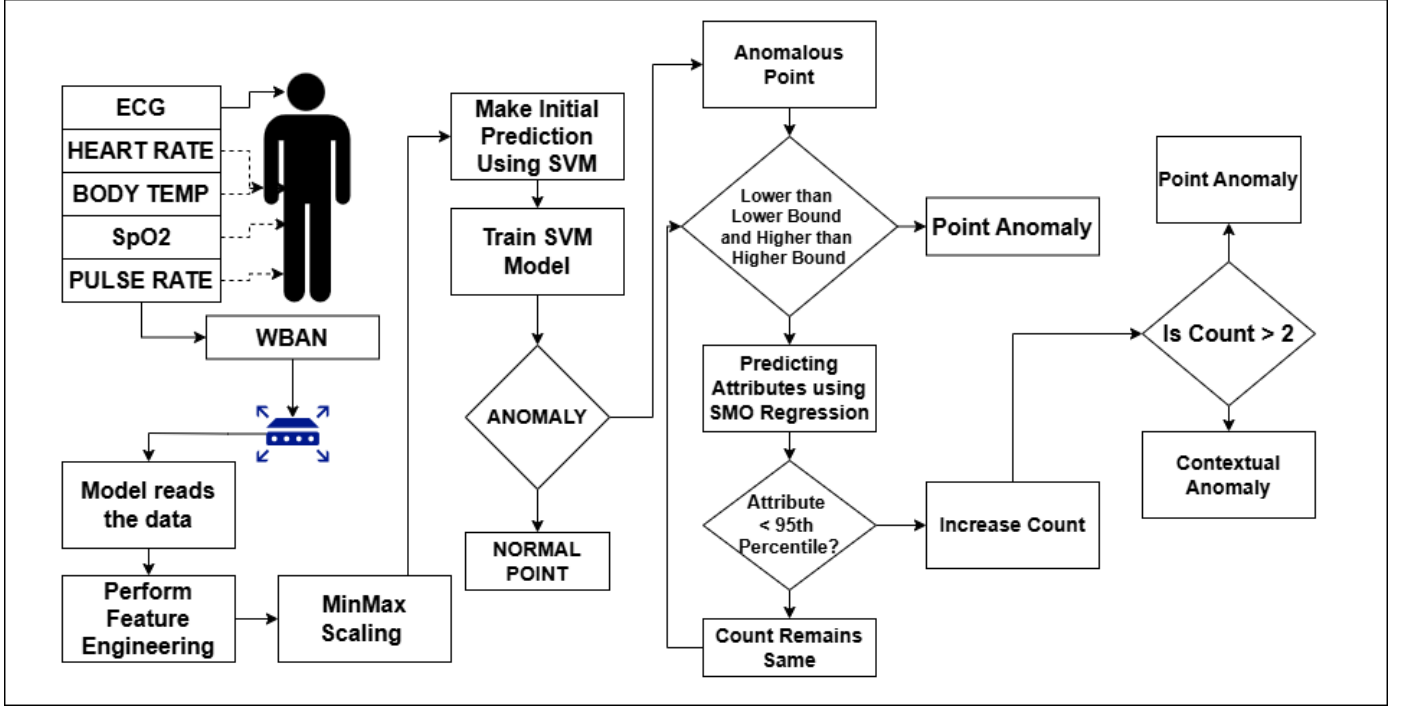


Fig. 2: System Architecture

algorithm utilizing Wireless Body Area Network (WBAN) data for personalized healthcare, Fig. 2. The system combines Support Vector Machines (SVM) and SMO regression (Sequential Minimal Optimization) models to identify the point and contextual anomalies in health data, which feature improved engineering techniques. Support Vector Regression (SVR) and Sequential Minimal Optimization (SMO) are closely related because SMO is a practical approach for resolving the optimization problem in SVR. By decreasing prediction errors within a margin, SVR (optimized via SMO) is used in anomaly detection to model expected behavior and identify data points that drastically vary from this pattern as anomalies. In the second stage, SMO regression is used to model and forecast the typical behavior of individual health measurements, including heart rate and ECG, using the context that other features give. Contextual anomalies, or situations in which health measurements substantially depart from anticipated patterns, are then found by analyzing the residuals from these predictions. The system adjusts to each person's health baseline by incorporating domain-specific thresholds and personalized residual analysis, guaranteeing that it can identify anomalies more accurately. This method offers better detection for early medical intervention and timely alarms, making it especially appropriate for personalized healthcare applications.

## II. BACKGROUND AND RELATED WORKS

The SMOReg, Gaussian Process, and Majority Voting (MV) algorithms are used in the paper [1] to design an anomaly detection system for human physiological data. The system computed dynamic error thresholds, forecasted sensor readings, and compared forecasts to real data to identify abnormalities. MV was utilized to verify real medical issues or false alarms, and anomalies were flagged based on surpassing the threshold. However, our model supports dynamic thresholds using machine learning techniques, which were not developed in this paper. In paper [2], the study presented a two-tier approach for efficient health monitoring in Wireless Body Area Networks (WBANs). The first tier involved eliminating unimportant health data to save energy for measurements with little deviation or likely faults. The second tier focused on anomaly detection at the Local Processing Unit (LPU), avoiding the need for cloud transmission, thus reducing latency and energy consumption. The paper [3] presented an anomaly detection approach for Medical Wireless Sensor Networks (WSNs) using real patient data from the Physionet database. The approach leveraged SVM to detect health anomalies. However, the approach relied heavily on static thresholds for anomaly detection, which may not account for dynamic patient conditions or environmental factors, leading to potential inaccuracies. In our model, to help define thresholds, we used

SMO regression, hence dynamic thresholds. Our model used various regression models to compare and get the best results. To learn about the models in detail, the paper [4] introduced a novel Logarithmic Kernel Function (LKF) for Support Vector Machines (SVMs) designed to improve performance with small datasets (10% and 30% of training samples). The LKF, with a Gaussian-like curve, outperforms traditional kernels (RBF, TANH) in classification and regression tasks, offering faster training times and lower support vector counts. In [5], the study provides a Markov Model-based approach for detecting anomalies in WBANs. A single-variate time series is used to discover anomalies by calculating the root mean square error (RMSE) between projected and actual values. Our model shows better results when using SMO Regression and SVM. To differentiate between real and false alarms, paper [6] proposes a sensor anomaly detection system for the healthcare industry that uses dynamic thresholds and SMO regression. The trials use artificial anomalies, which might not accurately reflect data from the real world. We employ real-world anomalies in our model. To improve detection rates and lower false alarms in big datasets, paper [7] has proposed research that suggests a hybrid anomaly detection model that combines K-means, DBSCAN, and Sequential Minimal Optimization (SMO). Choosing the correct parameters (such as the number of clusters and neighborhood radius) for K-means and DBSCAN can be challenging, which affects their performance. Our model performs better than this model, which will be discussed in upcoming sections. A hybrid anomaly detection approach that combines SMO classification with K-means clustering is proposed in the work [8]. Compared to previous models, it achieves 60.6% greater detection rates and 74.97% higher accuracy, greatly enhancing detection performance. The method lowers the false alarm probability (FAP) and has promise for real-time anomaly identification in big datasets in the future. Our approach investigates both point and contextual anomaly and outperforms this model.

### III. MODEL ARCHITECTURE AND DATASET

#### A. Dataset Used

Consisting of 72,000 rows of time-series data from 16 persons over 5 days, the SmartNet AI Labs gathered the dataset. Each individual's data was recorded thrice daily for five minutes, yielding 300 rows per session. Four people are defined as "patients" who display abnormal physiological patterns, and twelve are classified as "normal" in the dataset. Both contextual anomalies (shown by the "normal" labels) and point abnormalities (shown by the "column anomaly" labels) are captured in the dataset. This configuration provides a thorough dataset to assess the efficacy of anomaly detection models by simulating real-world situations where physiological signals

change due to different medical disorders. High-quality, real-time physiological sensors were used to collect the data: the ECG AD8232 for EKG readings, the DFRobot heart rate sensor, the MAX30102 for  $SpO_2$  and heart rate, and the MLX90614 for body temperature, Fig. 3. The participants were tracked by these Arduino-based sensors, which provided accurate physiological data for the anomaly detection model's training and testing. This dataset is perfect for evaluating the effectiveness of the suggested anomaly detection algorithm because it contains a wide range of physiological characteristics, both normal and abnormal.

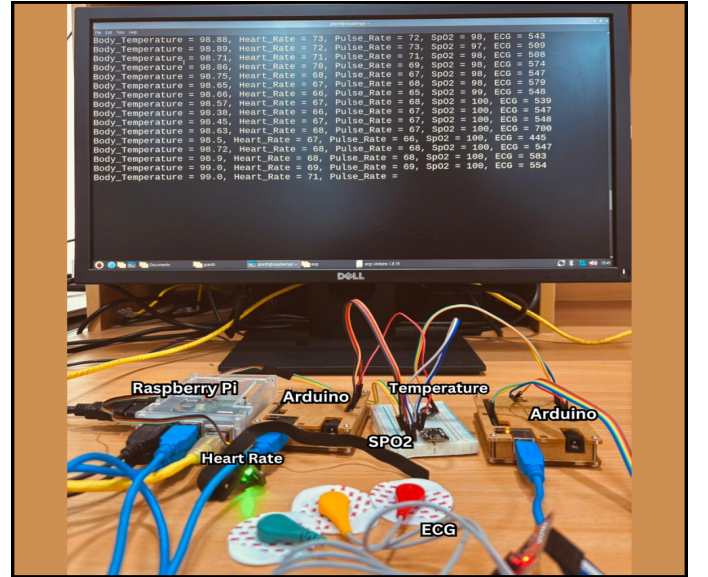


Fig. 3: IoT Test Bed Setup used to collect Data

#### B. Proposed Model

The suggested approach detects physiological data anomalies using two Support Vector Machines (SVM) stages. It uses feature engineering, scaling, and advanced regression approaches to identify anomalies as point or contextual, resulting in excellent accuracy and interpretability in anomaly identification.

1) *SVM for Identifying Anomalies:* An efficient supervised learning technique for binary classification is SVM. SVM allows for non-linear separability in high-dimensional feature spaces by modeling the boundary between normal and anomalous classes using a kernel function in the context of anomaly detection. Because of its capacity to manage intricate, non-linear relationships in the data, the Radial Basis Function (RBF) kernel was selected for this application. The SVM initially classifies Test samples as normal (class 0) or abnormal (class 1). In the RBF equation,  $K(x_i, x_j)$  is the kernel function between two data points  $x_i$  and  $x_j$ .  $\gamma$  is a parameter that defines

the influence of a single training example (kernel width).  $\|x_i - x_j\|^2$  is the squared Euclidean distance between  $x_i$  and  $x_j$ .

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2) \quad (1)$$

2) *Feature engineering and scaling*: To ensure the characteristics fall within the same range, Min-Max scaling normalizes them before applying the SVM. This preprocessing step enhances the convergence and performance of the model. The designed elements, such as rolling statistics for ECG, average heart rate, and heart rate difference, provide the input data with context, improving the model's capacity to distinguish between typical and abnormal patterns. The equation calculates the rolling standard deviation of ECG signals for an individual  $v$  at time  $t$ , denoted as  $\text{ECG\_RStd}_{v,t}$ . Here,  $w$  is the window size,  $\text{ECG}_{v,t-k}$  represents the ECG value  $k$  steps before time  $t$ , and  $\text{ECG\_RMean}_{v,t}$  is the rolling mean of the ECG signal over the same window.

$$\text{Avg\_Heart\_Rate}_i = \frac{\text{Heart\_Rate1}_i + \text{Heart\_Rate2}_i}{2} \quad (2)$$

$$\text{ECG\_RMean}_{v,t} = \frac{1}{w} \sum_{k=0}^{w-1} \text{ECG}_{v,t-k} \quad (3)$$

$$\text{ECG\_RStd}_{v,t} = \sqrt{\frac{1}{w} \sum_{k=0}^{w-1} (\text{ECG}_{v,t-k} - \text{ECG\_RMean}_{v,t})^2} \quad (4)$$

3) *Contextual Anomaly Detection using SVR*: The second step involves training Support Vector Regression (SVR) models for every feature to determine baseline residuals. These residuals serve as detection criteria for contextual anomalies. Fig. 4 and Fig. 5 capture the Contextual Anomalies of WBAN features. The SVR models forecast expected feature values using all other features as input. Test samples beyond these levels are marked as contextual anomalies highlighting multi-feature deviations.

In this model, feature engineering is essential. Three designed features were obtained: ECG Rolling Statistics (mean and standard deviation), Heart Rate Difference, and Average Heart Rate. These characteristics aid in the detection of patterns in physiological data, the smoothing of sensor noise, and the capturing of inter-sensor differences. In Algorithm I, the 95th percentile of residuals seen during training served as the baseline residual threshold. A record is marked as a contextual anomaly if residuals for two or more features exceed the threshold. This method, which separates contextual anomalies from isolated sensor errors or noise, is based on physiological interdependence; in actual health emergencies, numerous

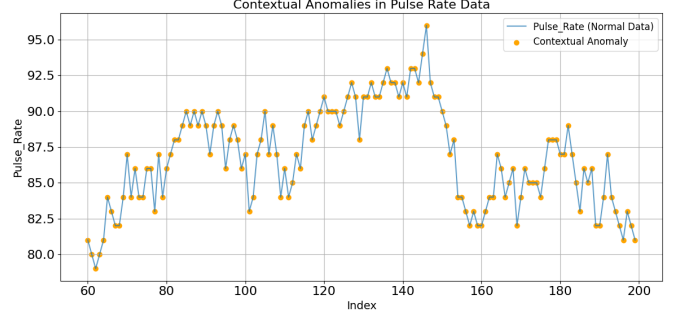


Fig. 4: Contextual Anomaly for Pulse Rate

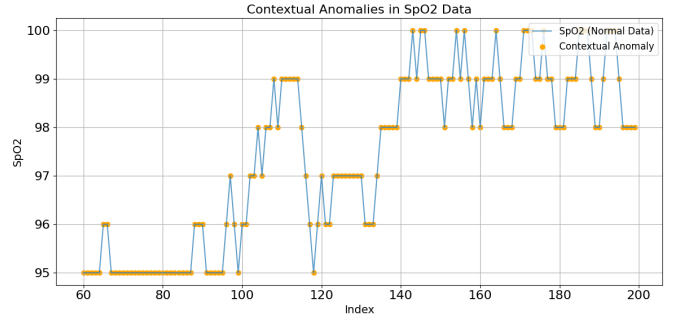


Fig. 5: Contextual Anomaly for SpO2

features diverge due to coupled biological responses. This two-step procedure combines machine learning and domain expertise, allowing for the reliable identification of discrete anomalies and intricate patterns suggestive of possible health problems.

---

**Algorithm 1** Anomaly Detection Using SVM and SVR Models

---

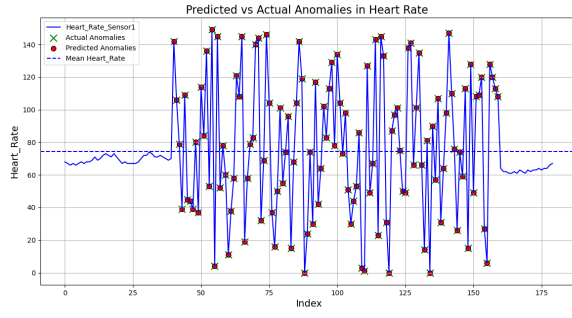
```

1: procedure ANOMALYDETECTION(input)
2:   1: Train SVM Model
3:   SVM_model  $\leftarrow$  TrainSVM( $X_{\text{train\_scaled}}, Y_{\text{train}}$ )
4:   SVM_predictions  $\leftarrow$  SVM_model( $X_{\text{test\_scaled}}$ )
5:   2: Compute Baseline Residuals with SVR
6:    $e_{i,f} \leftarrow |\text{SVR}_f(X_{\text{train\_scaled}}, -f) - X_{\text{train\_scaled},i,f}|$ 
7:   baseline_residual_f  $\leftarrow$  Percentile95( $\{e_{i,f}\}$ )
8:   3: Detect Anomalies
9:    $f_i \leftarrow \text{SVR}_f(X_{\text{test\_scaled}}, -f)$ 
10:   $e_{i,f} \leftarrow |f_i - X_{\text{test\_scaled},i,f}|$ 
11:  if  $e_{i,f} > \text{baseline\_residual}_f$  then
12:    Anomalyi  $\leftarrow$  1
13:  else if  $\sum_f \mathbf{1}_{\{e_{i,f} > \text{baseline\_residual}_f\}} \geq 2$  then
14:    Anomalyi  $\leftarrow$  2 // Contextual anomaly
15:  else
16:    Anomalyi  $\leftarrow$  0 // No anomaly
17:  return Anomaly(output)

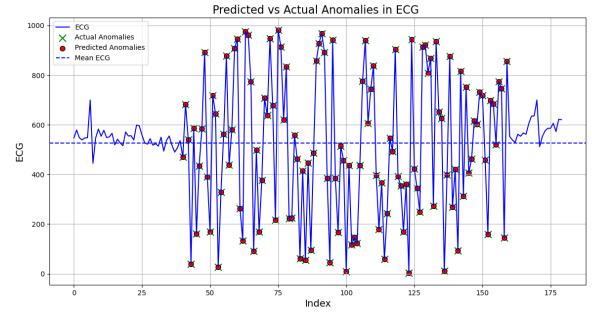
```

---

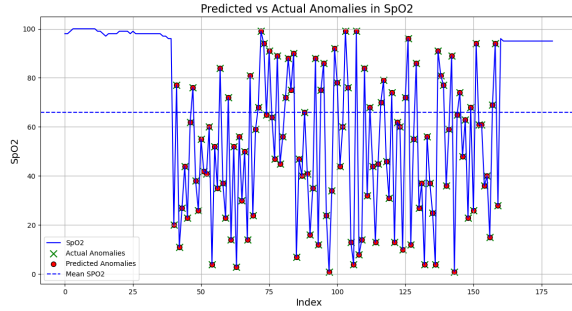




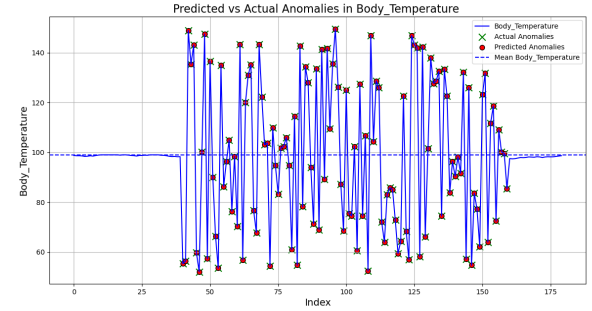
(a) Heart Rate



(b) ECG



(c) SPO2



(d) Body Temperature

Fig. 6: Predicted vs Original Anomalies

#### IV. EXPERIMENTAL RESULTS AND DISCUSS

The suggested anomaly detection system integrates machine learning approaches with domain expertise to effectively detect point and contextual anomalies. Fig. 6 shows the predicted anomalies and original anomalies of the features collected from WBANs. We can see that most predicted anomalies are overly on the original anomalies, showing that our model works well. Fig. 7 demonstrates that while the raw ECG feature has limited utility in anomaly detection, the derived features—ECG Rolling Mean and ECG Rolling Standard Deviation—offer more excellent value. By incorporating rolling statistics, the model can effectively analyze trends and variations over time in the time series data, improving its ability to detect anomalies linked to physiological changes. This emphasizes the significance of feature engineering in enhancing model performance.

Overall Accuracy measures the proportion of correctly identified anomalies (both point and contextual) out of all the samples. Point accuracy focuses on the fraction of correctly detected point anomalies relative to the total number of actual point anomalies. Contextual accuracy calculates the ratio of correctly detected contextual anomalies to the total number of actual contextual anomalies. These metrics help assess the

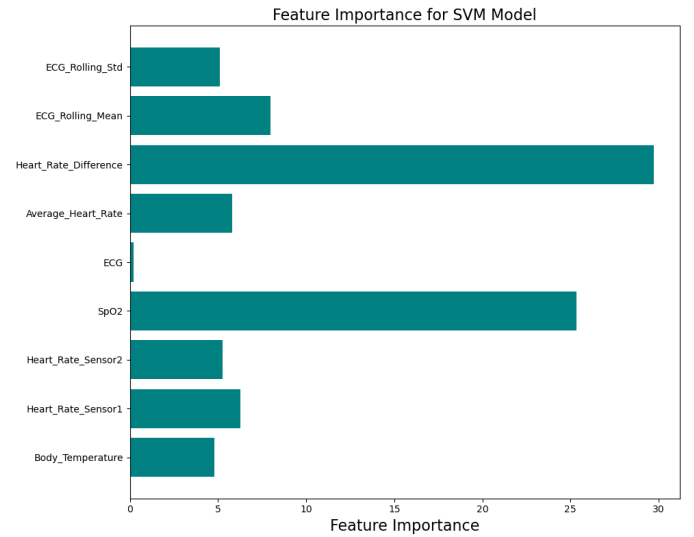


Fig. 7: Feature Importance for SVM

model's performance in detecting different types of anomalies.

$$\text{Overall\_Accuracy} = \frac{\text{Correct Anomaly Predictions}}{\text{Total Samples}} \quad (5)$$

$$\text{Point\_Accuracy} = \frac{\text{Correct Point Anomalies}}{\text{Total Point Anomalies}} \quad (6)$$

$$\text{Contextual\_Accuracy} = \frac{\text{Correct Contextual Anomalies}}{\text{Total Contextual Anomalies}} \quad (7)$$

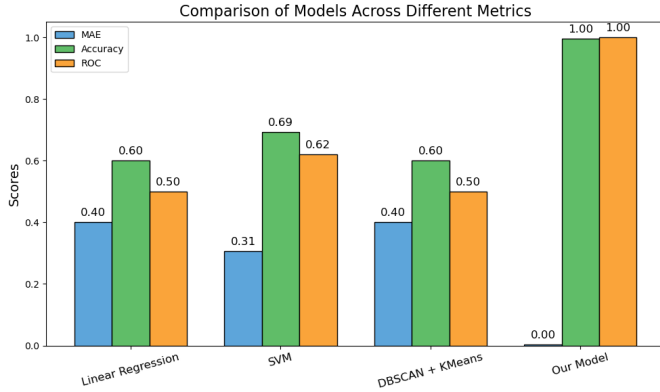


Fig. 8: Comparison of Accuracy, MAE and ROC

Our model performs exceptionally well on various evaluation metrics, surpassing baseline techniques, Fig. 8. With an overall anomaly detection accuracy of **99.59%**, the model significantly outperforms DBSCAN with KMeans (60.01%), SVM (69.29%), and linear regression (60.00%). Significantly lower than baseline models like SVM (0.3071) and DBSCAN with KMeans (0.3999), the mean absolute error (MAE) is 0.0041. With an overall accuracy of 100% across 18,000 test samples, the precision, recall, and F1-score for both the anomalous and regular classes also surpass 99%. With an accuracy of **99.61%** for point anomalies and **99.97%** for contextual anomalies, the model demonstrates exceptional anomaly identification granularity and robustness in detecting localized and associated patterns.

## CONCLUSION

The two-stage anomaly detection methodology presented in this paper uses Support Vector Machines (SVM) and Support Vector Regression (SVR) to detect contextual and point anomalies in physiological data. The suggested method successfully separates physiologically significant patterns from sensor noise by combining machine learning approaches with domain-specific thresholding, improving detection reliability. The combination of statistical and machine learning approaches and sophisticated feature engineering techniques like average heart rate computations, rolling statistics for ECG, and personalized thresholding based on individual medical histories are

important developments. These developments greatly increase the model's effectiveness in anomaly detection by allowing it to capture intricate relationships between physiological data. The proposed approach achieved detection accuracies of 99.61% and 99.97% for point and contextual anomalies, respectively, resulting in an overall anomaly detection accuracy of 99.59%. On the other hand, linear regression, DBSCAN with KMeans, and standalone SVM models achieved accuracies of 69.29%, 60.01%, and 60.00%, respectively. A mean absolute error of 0.0041 further demonstrates the remarkable precision attained by the suggested approach. The improved area under the curve (AUC) scores and receiver operating characteristic (ROC) curves further demonstrate its superior performance. The model, created as a scalable solution for real-time physiological health monitoring, may be applied to various healthcare applications because it incorporates customized thresholds, guaranteeing personalized detection based on distinct user profiles.

## ACKNOWLEDGMENT

The authors would like to thank BITS Bio-CyTiH Foundation and DST, Govt. of India for their support and funding for this research.

## REFERENCES

- [1] M. U. Harun Al Rasyid, F. Setiawan, I. U. Nadhori, A. Sudarsonc, and N. Tamami, "Anomalous data detection in wban measurements," in *2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC)*, 2018, pp. 303–309.
- [2] S. Jain, P. Jain, P. Upadhyay, J. Moualeu, and A. Srivastava, "An energy efficient health monitoring approach with wireless body area networks," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 3, p. 1–22, Apr 2022.
- [3] O. Salem, A. Guerassimov, A. Mehaoua, A. Marcus, and B. Furht, "Anomaly detection in medical wireless sensor networks using SVM and linear regression models," *Int. J. E-health Med. Commun.*, vol. 5, no. 1, pp. 20–45, Jan. 2014.
- [4] B. Hicdurmaz, N. Calik, and S. Ustebay, "Gauss-like logarithmic kernel function to improve the performance of kernel machines on the small datasets," *Pattern Recognition Letters*, vol. 179, pp. 178–184, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S016786552400014X>
- [5] M. U. Harun Al Rasyid, I. U. Nadhori, I. Syarif, I. Winarno, F. Furoida, and A. Amrullah, "Anomaly detection in wireless body area network using mahalanobis distance and sequential minimal optimization regression," in *2021 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2021, pp. 64–69.
- [6] S. A. Haque, M. Rahman, and S. M. Aziz, "Sensor anomaly detection in wireless sensor networks for healthcare," *Sensors (Basel)*, vol. 15, no. 4, pp. 8764–8786, Apr. 2015.
- [7] U. Rashid, M. F. Saleem, S. Rasool, A. Abdullah, H. Mustafa, and A. Iqbal, "Anomaly detection using clustering (k-means with dbscan) and smo," *Journal of Computing and Biomedical Informatics*, vol. 7, no. 02, Sep. 2024. [Online]. Available: <https://jcibi.org/index.php/Main/article/view/598>
- [8] S. Gadal, R. Mokhtar, M. Abdelhaq, R. Alsaqour, E. S. Ali, and R. Saeed, "Machine learning-based anomaly detection using k-mean array and sequential minimal optimization," *Electronics*, vol. 11, no. 14, 2022. [Online]. Available: <https://www.mdpi.com/2079-9292/11/14/2158>