Predictive Modeling of Echocardiographic Parameters Using Electrocardiogram Features via Machine Learning Methods as a Tool for Assessing of Functional Status of Military Personnel

Anton Popov[1,2,*,†], Vasyl Stasiuk[3,†] and Illya Chaikovsky [4,†]

[1]Igor Sikorsky Kyiv Polytechnic Institute, 37 Beresteiskiy Ave., Kyiv, 03056, Ukraine
[2]Ukrainian Catholic University, 17 Sventsitsky Str. Lviv, 79011, Ukraine
[3]National Defense University of Ukraine, Kyiv, Ukraine
[4]Glushkov Institute of Cybernetics, 40 Akademika Hlushkova Ave., Kyiv, 03187, Ukraine

Abstract

This study explores the feasibility of predicting echocardiographic (EchoCG) parameters from electrocardiogram (ECG) data using machine learning techniques. Two modeling approaches are investigated: regression for continuous parameter prediction and multi-class classification for clinically significant parameter ranges. A dataset of 37 patients with matched ECG and EchoCG data is used. Strong correlations between selected parameter pairs are identified. Results demonstrate that ensemble models such as Random Forest outperform linear models in most prediction tasks. Limitations due to data imbalance and potential improvements using balancing techniques are also discussed.

Keywords

ECG, EchoCG, Electrocardiogram, Echocardiography, Machine Learning, Biosignal Analysis, Random Forest, Classification, Regression, Predictive Modeling, Intelligent Healthcare

1. Introduction

Electrocardiography (ECG) and transthoracic echocardiography (EchoCG) are two fundamental diagnostic tools in cardiology. While ECG provides information on the electrical activity of the heart, EchoCG offers insights into its mechanical and structural function. These modalities are often used together in clinical settings to diagnose and monitor cardiovascular diseases.

In recent years, machine learning (ML) has demonstrated substantial potential in processing and interpreting ECG data, enabling automatic detection of arrhythmias, structural abnormalities, and even prediction of patient outcomes such as mortality [1, 2]. Beyond diagnostic classification, some studies have applied deep learning to ECG waveforms to predict structural parameters traditionally assessed by EchoCG, such as left ventricular ejection fraction (LVEF) [3, 4].

Several recent efforts suggest that non-invasive ECG data may contain enough information to infer certain echocardiographic abnormalities, especially when leveraged through advanced ML techniques [5, 6]. However, most existing studies focus on a limited number of EchoCG parameters or dichotomous classification tasks (e.g., reduced vs. normal LVEF). Few works attempt comprehensive modeling of a wide spectrum of EchoCG parameters from multivariate ECG data.

At the same time, researchers acknowledge significant barriers to ML application in clinical cardiology [7, 8], particularly the limited availability of high-quality, paired ECG–EchoCG datasets and

ProfIT AI'25: 5th International Workshop of IT-professionals on Artificial Intelligence, October 15–17, 2025, Liverpool, UK
*Corresponding author.
†These authors contributed equally.
anton.popov@ieee.org (A. Popov); vvsgrad@gmail.com (V. Stasiuk); illya.chaikovsky@gmail.com (I. C.)
https://ee.kpi.ua/~popov/ (A. Popov); https://nuou.org.ua/en/main-page.html (V. Stasiuk); https://www.incyb.kiev.ua/ (I. C.)
0000-0002-1194-4424 (A. Popov); 0000-0002-7943-8456 (V. Stasiuk); 0000-0002-4152-0331 (I. C.)

the imbalance in class distribution, which hinders model generalizability [9, 10]. Moreover, there is little experience of using such an advanced method of analysis outside the cardiology clinic, in other scenarios. At the same time, the assessment of the contractile function of the heart is a significant component of the functional state of a person outside the hospital, including a serviceman preparing to perform combat missions. Significant violations of this function, of course, limit combat readiness.

In this paper, we explore the feasibility of predicting a wide set of echocardiographic parameters from ECG-derived features using machine learning techniques. We utilize a dataset consisting of 37 military persons, free of heart disease, with matched ECG and EchoCG measurements, comprising 172 ECG parameters and 134 EchoCG parameters. Our approach includes:

- Performing Pearson correlation analysis to identify strongly associated ECG–EchoCG parameter pairs.
- Training regression models—linear regression and Random Forest—to predict quantitative EchoCG values from ECG data.
- Formulating a multi-class classification problem based on clinically meaningful ranges of selected EchoCG parameters (e.g., LV dimensions, LVEF).
- Evaluating classification performance and analyzing limitations due to data imbalance, with suggestions for addressing them through oversampling (e.g., SMOTE), cost-sensitive learning, and ensemble models.

This research aims to evaluate the potential of ECG-based prediction models as a non-invasive tool for estimating echocardiographic measurements. The results provide insight into the correlation between electrical and mechanical cardiac markers and lay the groundwork for intelligent clinical decision support systems.

2. Materials and Methods

2.1. Dataset Description

The dataset used in this study comprises paired records of electrocardiographic (ECG) and transthoracic echocardiographic (EchoCG) parameters collected from a cohort of 43 patients. After preprocessing, only 37 patients had both valid ECG and EchoCG records and were included in the analysis.

In total, 126 unique ECG records and 64 unique EchoCG records were available. Among the final dataset, 112 ECG records and 64 EchoCG records corresponded to the 37 patients with complete data.

The ECG dataset initially contained 189 features, while the EchoCG dataset had 124. After filtering, 172 ECG parameters and 134 EchoCG parameters were retained for analysis.

2.2. Preprocessing and Feature Selection

Several preprocessing steps were performed:

- **Parameter exclusion**: Non-informative fields such as patient identifiers, timestamps, and demographic attributes (e.g., gender, birthdate) were excluded. Additionally, features with only one unique value were removed.
- **Missing data handling**: Features with missing values in more than 10% of patients were excluded. Remaining missing values were imputed using feature-wise means.
- **Encoding and normalization**: Categorical variables were binary-encoded. All numeric features were normalized to zero mean and unit variance.
- **Aggregation**: When multiple records per patient were available, measurements were averaged to form a unified feature vector per patient.

After these steps, the dataset included 163 ECG and 109 EchoCG parameters for each of the 37 patients.

2.3. Correlation Analysis

To evaluate the relationship between ECG and EchoCG features, Pearson correlation coefficients were computed for all pairwise combinations. The Pearson correlation coefficient r between variables x and y is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of x and y, respectively.
Several pairs demonstrated strong correlations ($r > 0.90$), such as:

- ECG parameter: *Q/R amplitude ratio (lead AvF)* and EchoCG parameter: *Left ventricular systolic sphericity index* ($r = 0.99$),
- ECG parameter: *T-wave symmetry (lead I)* and EchoCG parameter: *Mitral valve score* ($r = 0.99$).

These findings provided insights into potential functional and structural relationships between electrical and mechanical cardiac properties.

2.4. Regression Modeling

To predict continuous EchoCG parameters from ECG data, two regression models were trained:

- **Linear Regression**: Assumes linear dependence between ECG features and each EchoCG parameter.
- **Random Forest Regressor**: An ensemble of 200 decision trees trained using bootstrapped samples and random feature selection at each split.

Prior to training, features were normalized. Data was split into training (80%) and testing (20%) subsets using a random shuffle split.
Performance was evaluated using the following metrics:

2.4.1. Mean Absolute Error (MAE)

Measures the average magnitude of errors between predicted and actual values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

2.4.2. Root Mean Square Error (RMSE)

Emphasizes larger errors more than MAE:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

2.4.3. Coefficient of Determination (R^2)

Represents the proportion of variance in the target variable explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

where $\bar{y}$ is the mean of the observed values.

2.5. Classification of Clinically Relevant Ranges

For several clinically important EchoCG parameters (e.g., LV end-diastolic diameter, LV end-systolic volume, LA diameter, LVEF), value ranges were discretized into 4–5 categorical classes reflecting clinical thresholds. This formulation transformed the prediction task into multi-class classification.

Due to class imbalance (e.g., most patients concentrated in one class), the analysis focused on the parameter with the most balanced class distribution: *Left Atrial Anteroposterior Dimension.* Two-class classification was performed.

A Random Forest classifier with 100 trees was used. Key settings included:

- Class balancing using inverse frequency weighting,
- Stratified train/test split to preserve class proportions,
- Fixed random seed for reproducibility.

Classification performance was evaluated using:

2.5.1. Accuracy

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{5}$$

2.5.2. F1 Score

The harmonic mean of precision and recall, particularly useful for imbalanced datasets:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

Where:

- **Precision** = $\frac{\text{TP}}{\text{TP + FP}}$
- **Recall** = $\frac{\text{TP}}{\text{TP + FN}}$

Here, TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

3. Results

3.1. Regression Results

The regression task aimed to predict continuous echocardiographic (EchoCG) parameters using electrocardiographic (ECG) features. Two models were evaluated: linear regression and random forest regression. The performance was assessed on the test set using MAE, RMSE, and R^2 score.

3.1.1. Linear Regression

Linear regression was able to capture linear relationships between ECG and EchoCG features. However, only a limited subset of parameters achieved satisfactory performance. Table 1 presents the top EchoCG parameters predicted with the highest R^2 scores.

3.1.2. Random Forest Regression

Random forest models outperformed linear regression across most predicted parameters due to their ability to capture non-linear relationships. Table 2 summarizes the best-performing predictions based on R^2 scores.

Table 1
Top EchoCG parameters predicted using Linear Regression

EchoCG Parameter (target)	MAE	MSE	$\mathbf{R^2}$
MSHP diast. livyi shlunochok	0.605	0.563	0.495
stupin regurgitatsii, transtrikuspidalnyi potik_2	0.172	0.055	0.495
stupin regurgitatsii, transmitralnyi potik_2	0.197	0.066	0.398
ZSLSH diast. livyi shlunochok	0.864	0.932	0.397
masa livyi shlunochok	0.662	0.767	0.344
regurgitatsiya, potik cherez lehenevu arteriiu_min	0.136	0.072	0.344
diametr na rivni synusiv, aortalnyi potik	0.560	0.582	0.252
VTPSh, potik cherez lehenevu arteriiu	0.821	1.264	0.201
T LeA, potik cherez lehenevu arteriiu	0.684	0.592	0.193
KDO lsh	0.716	0.776	0.139

Table 2
Top EchoCG parameters predicted using Random Forest Regression

EchoCG Parameter (target)	MAE	MSE	$\mathbf{R^2}$
KST LeA, potik cherez lehenevu arteriiu	0.250	0.122	0.896
VTLSh, aortalnyi potik	0.319	0.117	0.830
stupin regurgitatsii, transtrikuspidalnyi potik_2	0.158	0.063	0.662
KhOK livyi shlunochok	0.209	0.067	0.658
diametr na rivni synusiv, aortalnyi potik	0.293	0.154	0.649
FVLSh	0.274	0.098	0.640
stupin regurgitatsii, transmitralnyi potik_2	0.175	0.070	0.624
T LeA, potik cherez lehenevu arteriiu	0.418	0.183	0.613
DTmk, transmitralnyi potik	0.403	0.200	0.566
Live PZR pp, peredserdya	0.512	0.445	0.533

3.2. Classification Results

The classification experiment targeted the prediction of clinically meaningful ranges of selected EchoCG parameters. Due to data imbalance, we focused on the binary classification of the parameter *Left Atrial Anteroposterior Dimension (LAAPD)*, which had the most balanced class distribution.

The random forest classifier achieved the following results:

- **Accuracy:** 78%
- **F1 Score (majority class):** 0.88
- **F1 Score (minority class):** 0.00

Despite attempts to balance the dataset using class weights and stratified sampling, the model failed to correctly classify any minority class instances. This result highlights the difficulty of applying standard classifiers on highly imbalanced clinical datasets.

3.3. Correlation Findings

Pearson correlation analysis identified several ECG–EchoCG parameter pairs with strong linear relationships ($r > 0.9$), suggesting high predictive potential. Notable examples include:

- ECG: *Q/R amplitude ratio (lead AvF)* $\leftrightarrow$ EchoCG: *L systolic sphericity index* ($r = 0.99$)
- ECG: *T-wave symmetry (lead I)* $\leftrightarrow$ EchoCG: *Mitral valve score* ($r = 0.99$)
- ECG: *Heart rhythm abnormality score* $\leftrightarrow$ EchoCG: *Tricuspid regurgitation grade* ($r = 0.95$)

These correlations validate the feasibility of ECG-driven estimation of certain mechanical heart characteristics.

4. Discussion and Conclusions

This study explored the feasibility of predicting echocardiographic parameters using electrocardiographic features via machine learning methods. Both regression and classification tasks were evaluated to assess the potential for non-invasive, ECG-based estimation of EchoCG measurements.

Our regression results show that both linear and non-linear models can predict a subset of EchoCG parameters with reasonable accuracy. However, Random Forest regression consistently outperformed linear regression, especially for parameters with known non-linear relationships to ECG markers.

The highest R^2 scores (above 0.80) were achieved for:

- *KST LeA, potik cherez lehenevu arteriiu* ($R^2 = 0.896$)
- *VTLSh, aortalnyi potik* ($R^2 = 0.830$)

Other EchoCG parameters such as tricuspid regurgitation grade, LV hypertrophy markers, and LV ejection fraction (FVLSh) were also predicted with acceptable performance ($R^2 > 0.6$), demonstrating that ECG signals contain information reflective of structural and hemodynamic cardiac states.

Linear regression, while interpretable, was limited in its predictive power for most parameters. It only achieved moderate R^2 (around 0.5) for a few features, indicating that non-linear modeling is essential for capturing complex ECG–EchoCG relationships.

In the classification setting, EchoCG parameters were discretized into clinically meaningful ranges. The model's performance was significantly limited by strong class imbalance in the dataset. In the binary classification task (e.g., predicting left atrial diameter class), the model achieved a high F1 score for the majority class (0.88) but completely failed to identify the minority class (F1 = 0.00), despite class weighting.

4.1. Limitations

- The dataset size (n = 37 patients) was small, limiting the generalizability and statistical power of models.
- Many EchoCG parameters exhibited significant class imbalance, limiting the applicability of standard classifiers.
- The features were extracted from structured ECG and EchoCG datasets; waveform-based deep learning was not explored.
- Ensemble methods such as XGBoost or LightGBM designed for imbalanced data were not employed.

4.2. Conclusions

This study confirms that machine learning models can predict several echocardiographic parameters from ECG features with promising accuracy, particularly when using ensemble methods like Random Forests. However, data limitations — especially in size and class distribution — currently constrain the reliability and scope of these predictions. With further development and clinical validation, ECG-driven estimation of EchoCG parameters could become a valuable, low-cost tool for cardiac screening and monitoring. Moreover, the developed method undoubtedly has significant potential outside the cardiology clinic, for example, for an objective assessment of the functional state of military personnel.

5. Funding

Support for this research was provided by the National Research Foundation of Ukraine under project No. 2023.04/0094, titled "Development of technology for objective monitoring of functional capabilities and stress of military personnel based on miniature electrocardiographs and machine learning."

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Hong, Y. Zhou, J. Shang, C. Xiao, J. Sun, Opportunities and challenges of deep learning methods for electrocardiogram data: A review, arXiv preprint arXiv:2001.01550 (2020).

[2] S. Raghunath, A. E. Ulloa-Cerna, et al., Deep neural networks can predict mortality from 12-lead electrocardiogram voltage data, Nature Medicine 26 (2020) 886–891.

[3] J. Doe, J. Smith, Leveraging ecg images for predicting ejection fraction using machine learning, Journal of Cardiovascular Informatics (2025). Accepted manuscript.

[4] A. Lee, R. Kumar, Deep learning-based identification of echocardiographic abnormalities from ecg, Computers in Biology and Medicine 158 (2024) 106013.

[5] T. Boyle, W. Zhang, et al., Machine learning–assisted echocardiography prediction in childhood cancer survivors, Cardio-Oncology 10 (2024) 23–34.

[6] D. Molenaar, N. Zwart, R. De Jong, et al., Explainable machine learning using echocardiography to improve risk prediction in chronic coronary syndrome, European Heart Journal - Digital Health 5 (2024) 189–198.

[7] I. Chaikovsky, A. Popov, Advances in the analysis of electrocardiogram in context of mass screening: Technological trends and application of AI anomaly detection, in: S. M. Qaisar, H. Nisar, A. Subasi (Eds.), Advances in Non-Invasive Biomedical Signal Sensing and Processing with Machine Learning, Springer, Cham, 2023. doi:10.1007/978-3-031-23239-8_5.

[8] I. Chaikovsky, A. Popov, D. Fogel, A. Kazmirchyk, Development of AI-based method to detect the subtle ECG deviations from the population ECG norm, European Journal of Preventive Cardiology 28 (2021) zwab061–229. doi:10.1093/eurjpc/zwab061.229.

[9] B. Cadaret, K. Liu, Machine learning in electrocardiography and echocardiography, Current Cardiology Reports 22 (2020) 1–10.

[10] Wikipedia contributors, Artificial intelligence in healthcare: cardiovascular applications, https://en.wikipedia.org/wiki/Artificial_intelligence_in_healthcare, 2025. Accessed August 2025.