Arrhythmia Classification from ECG Signals Using Transformers and Data Balancing Techniques

Jaunė Malūkaitė, Jolita Bernatavičienė, Povilas Treigys

Vilnius University, Institute of Data Science and Digital Technologies, Akademijos str. 4 Vilnius
jaune.malukaite@mif.stud.vu.lt, jolita.bernataviciene@mif.vu.lt, povilas.treigys@mif.vu.lt

Abstract. While many arrhythmias pose minimal threat, certain heart rhythm irregularities elevate the potential for stroke or heart failure. The complexity arises particularly with the supraventricular premature heartbeat which has a resemblance to a normal beat and occurs infrequently. Consequently, this research proposes a data balancing and classification technique that enhances the accuracy of identifying mentioned hard-to-classify heartbeats while maintaining robust metrics for other classes. The study introduces a deep learning framework combined with a multi-head attention transformer, for balancing – under-sampling and synthetic minority oversampling are used. To evaluate the proposed model, various experiments based on real data were conducted. The results were compared with an existing model used in chest belt heartbeat monitoring, and the results show that the transformer model achieved better performance for supraventricular premature heartbeats, at the same time reaching high overall and per-class metrics.

Keywords: ECG signals, Classification, Deep Learning, Transformer, Focal Loss, Data Balancing Techniques, Heartbeats.

1 Introduction

According to the Lithuanian Institute of Hygiene, more than 22.5 thousand people in Lithuania died in 2022 due to diseases of the circulatory system, accounting for 53 % of all deaths in the country. International data also show that Lithuania's cardiovascular mortality rates are well above the European Union (EU) average and among the highest in the EU. While arrhythmias can be detected from electrocardiograms, the process is time-consuming and prone to errors even among experts. This underscores the significance of automated electrocardiogram analysis. Automated classification of

DOI: https://doi.org/10.15388/LMITT.2024.13

arrhythmias can alleviate the challenging daily workload for medical professionals and facilitate earlier identification of cardiac disorders in patients. Therefore, patients can receive appropriate treatment strategies earlier.

The most common classes of heartbeats analysed in studies are three or four – in this research three classes are chosen, namely supraventricular premature heartbeats (S), normal heartbeats (N) and ventricular premature contraction (V). For the classification of these heartbeats, different machine learning and deep learning models can be used. Approaches such as support vector machine, logistic regression, k-nearest neighbours, and random forest have been used in scientific literature and have demonstrated promising outcomes [3]. However, it has been noted that heart rate classification techniques, which depend on manually extracted features, frequently struggle to discern abstract relationships within the data. Therefore, there is an increasing number of research articles emphasizing the significance of using deep learning methodologies for this purpose [5]. Deep learning transformers have also become increasingly important in recent years, as they have attention mechanisms that give more weight to more important elements in the input sequence. In the arrhythmia classification task, the transformer relies on an attention mechanism and uses the electrocardiogram segments as input to capture global dependencies of signal values [5]. Researchers Rui Hu, Jie Chen, and Li Zhou propose a transformer and neural network architecture wherein a segmented one-dimensional electrocardiogram sequence serves as the input, undergoing multiple one-dimensional convolutional layers. The encoders within the transformer are constructed by iteratively stacking layers with identical structures containing a multi-head self-attention module and a feed-forward network featuring a single hidden layer [2]. The transformer exhibits versatility because it is adaptable not only to convolutional architectures but also to recurrent neural networks. Given the ability of recurrent neural networks to comprehend heart rhythm characteristics, employing a recurrent neural network-based sequence-to-sequence approach could prove advantageous in addressing cardiac classification challenges [4].

The widespread applicability of transformers is evident, hence, in this study, an architecture comprising deep neural networks and transformers is proposed. A significant challenge lies in achieving robust classification

accuracy when using patient data in the test dataset that was not included in the training dataset. Given the anatomical variations among individuals, models tend to emphasize these distinctions over differences in heartbeats themselves. Furthermore, data imbalance poses a recurring complication, as normal heartbeats are disproportionately represented compared to S or V heartbeats, leading to inflated overall metric values primarily driven by the abundance of N class data. Therefore, different data balancing methods are used to overcome this problem [1].

In this research, we seek to improve the supraventricular premature heartbeat classification recall metric by using a transformer model. Furthermore, the impact of different balancing techniques on classification metrics is analysed to find whether data balancing improves metrics. The rest of this paper is organized as follows. In Section 2, information about real data used in the research and its processing is provided. The proposed methodology is discussed in Section 3, while in Section 4 the experimental results are compared and presented. Finally, conclusions are drawn in Section 5.

2 Data

This study utilises data collected from a chest belt for heartbeat monitoring created by Zive company. The dataset consists of 1086 recordings from 102 patients, each lasting 10 minutes. Each recording is subsequently segmented into individual signals. Across the dataset, there are 730860 N heartbeats, 6550 S heartbeats, and 17463 V heartbeats. For additional data combinations and experiments, data from the PhysioNet MIT-BIH Arrhythmia Database is utilised. In the MIT-BIH dataset, which is also used for chest baseline CNN model training, there are 88349 N, 2668 S, and 6783 V heartbeats. While comparing the two datasets, the Zive dataset has more occurrences, especially N heartbeats.

The R peaks of the signals are identified to divide the recordings into heartbeat segments. Following the detection of each R peak, the local minima from the left and the right sides (Q and S' peaks) are found. These peaks create a QRS complex. Additionally, as shown in Figure 1, supplementary parameters such as RRl (the number of signal values to the left closest R peak) and RRr (the number of signal values to the right closest R peak) are computed to ascertain the signal length. The signal is defined as 70 % of RRl values to the left and 70 % of RRr values to the right. A transformation to functional data is then used to standardize all segments to a length of 200.

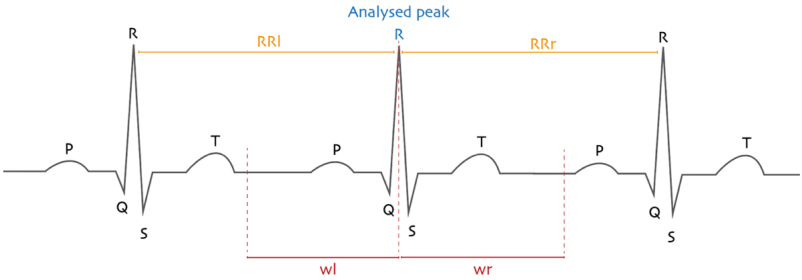

Figure 1. Additional parameters computed from signals. RRl – R peak from the left, RRr – R peak from the right, wl – 70 % of RRl, wr – 70 % of RRr.

Afterwards, data is divided into training, validation, and testing sets. Each set consists of different patient data that were divided into sets by hand to have similar class distributions in all datasets. In the datasets, all 200 signal values, additional derivative features, P, Q, R, S, P values and positions are used.

In the under-sampled dataset, N class occurrences are reduced to 100 thousand while in the SMOTE dataset, N class is reduced to 100 thousand, S class synthetically increased to 20 thousand and V to 40 thousand occurrences. For experimental analysis, different class proportions were used but in further analysis, it was decided to use classes S and V with a balance of 1:2 to have closer to real-life occurrence distribution, in addition to promising first received results. SMOTE numbers are chosen not that large because the synthetical creation can introduce noise and inaccurately imitate heartbeat data. In the original training dataset, there are 380025 N, 4796 S and 11572 V signal segments.

3 Methodology

For data balancing, two different techniques are used: random under-sampling of class N signals and synthetic minority oversampling (SMOTE) [1] of S and V signals in the training set. Random oversampling is not used in this case as the number of S and V classes is very low, and experiments showed that the model tends to learn how to identify only one class with high metrics. SMOTE algorithm creates a new sample for each instance x_i using x_i and its k nearest neighbours in feature space, as defined in the 1

equation. In the equation, x'_i is a new example synthesised from the sample x_i and a randomly selected sample x_j from the nearest neighbours of x_i and λ is a random value from the interval [0, 1] [1].

$$x'_i = x_i + \lambda(x_j - x_i) \quad (1)$$

A custom Focal Loss function is defined and used in the model which is particularly effective for imbalanced datasets, such as those often found in medical diagnosis tasks. This loss function prioritizes challenging instances over simpler ones by adjusting the alpha values used in computations, thereby enhancing focus on harder-to-classify examples. Focal Loss is implemented by adding a modulating factor to the Cross-Entropy loss. In Formula 2, α is considered a weighing factor, p_i is the predicted probability, b is the logarithm base, n is the number of elements being predicted while γ is the focusing parameter. γ rescales the modulating factor such that the easy examples are down-weighted more than the hard ones, reducing their impact on the loss function.

$$Focal\ Loss = -\sum_{i=1}^{i=n} \alpha_i (i - p_i)^{\gamma} \log_b(p_i) \quad (2)$$

The proposed model in the research is created using *PyTorch Lightning* newest version used for streamlined model development and training. The model shown in Figure 2 starts with a multi-headed self-attention mechanism allowing the model to focus on different parts of the ECG signal simultaneously. That is why the input length must be dividable by the number of heads used in the attention mechanism. The mechanism is followed by feed-forward networks within each encoder block. Layer normalization and residual connections stabilize training and facilitate deeper networks. For the transformer encoder, encoder blocks are repeated. Finally, the output from the transformer encoder is passed through a final layer to produce predictions for the ECG signal classes. The decoder part is not used as the encoded signal does not require translation back into a signal for class predictions. The model adopts a 6-layer depth, and a batch size of 128 because it yields better outcomes to 32, 64 or 256. What is more, dropout is integrated to prevent model overfitting, alongside the utilization of a learning rate scheduler and early stopping mechanisms. Optimal epoch checkpoints are stored based on validation loss criteria.

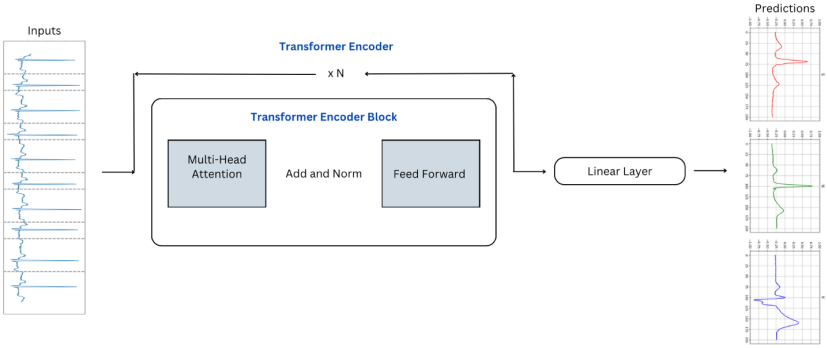

Figure 2. Proposed model architecture scheme that displays the overall model flow and transformer encoder elements.

Base line model is a convolutional neural network with 13 convolutional layers, each followed by batch normalization and activation layers. Batch normalization helps to stabilize and accelerate the training process, while activation layers introduce non-linearity to the network. Additionally, there are two dense layers, a global average pooling layer that helps to reduce the number of parameters. Finally, an output layer is added at the end of the model architecture.

4 Results

For different model and dataset results comparison recall is used as it shows the fraction of instances in a class that the model correctly classified out of all instances in that class. For overall metric calculation macro average recall is chosen to have the classes weighted equally as the amount of N occurrences would distort the results – the weighted metrics would be high even if individual class metrics would be low for S and V classes.

Comparing the model currently used in chest belts and the transformer model with different balance datasets, it is seen from Table 1 that the highest macro average recall for all classes combined is achieved using the transformer model and under-sampled dataset, the metric reaches 0.822 value. Because the research aims to increase class S recall while having N and V class high metrics, the transformer model with an under-sampled dataset achieves 0.570 recall for the S class, which is 0.092 higher than the model in chest belts. N and V class recall values are also high – 0.942 and

0.955 respectively. As for the transformer model with SMOTE dataset, it also achieves promising results, especially for the V class where recall rockets up to 0.983. Regarding the transformer model with original dataset, different parameters and architectures were tried but if one class metrics rise, the other two class metrics decrease.

Table 1. Model recall results using different balance datasets. The model whose results are aimed to be increased is also included for comparison.

Model	N recall	S recall	V recall	Macro avg. recall
Baseline model	**0.997**	0.478	0.934	0.803
Transformer model + under sampled dataset	0.942	**0.570**	0.955	**0.822**
Transformer model + SMOTE dataset	0.929	0.550	**0.983**	0.821
Transformer model + original dataset	0.932	0.559	0.877	0.789

The model used in chest belts has a high weighted average precision of 0.986, recall of 0.987, and f1-score of 0.986 as it learned well class N occurrences and predicts this class most of the time correctly. Because in the test dataset class N appears more often than class S, the overall metrics are high. As for the model with the highest macro average recall score, which is the transformer model used with an under-sampled dataset, the weighted average precision is 0.984, recall 0.939, and f1-score 0.959. In this case, the metrics are also high, while the predictions for the S class have improved.

5 Conclusions

In this paper, we proposed a transformer model which classifies ECG signals into three heartbeat classes. The model architecture and parameters are changed accordingly to experiments made using different balance datasets – under-sampled, oversampled using the SMOTE technique, and the original dataset. In the training, validation, and testing datasets, different patients were used to avoid bias and model learning features relevant to the individuals, not the differences in heartbeats. The best results are received using the proposed transformed model and an under-sampled dataset. The model achieved a macro average recall score of 82.2 %, while the accuracy for the S class, which is the hardest to classify, increased by 9.2 % comparing

the results with a baseline model. In the future, our focus is to increase even more class S metrics by trying out different class proportions in datasets, introducing noisy signals to understand how the model performs when the signals are not high quality, and using transformers together with other deep learning architectures.

6 Acknowledgements

We extend our gratitude to UAB Zive company for the collaboration in data collection, analysis, and comprehension. Additionally, we are thankful for the high-performance computing resources provided by the Information Technology Research Center of Vilnius University.

Research funded under the Programme "University Excellence Initiatives" of the Ministry of Education, Science and Sports of the Republic of Lithuania (Measure No. 12-001-01-01-01 "Improving the Research and Study Environment").

References

[1] J. Chen, J. Lalor, W. Liu, E. Druhl, E. Granillo, V. G. Vimalananda, and H. Yu. Detecting hypoglycemia incidents reported in patients' secure messages: using cost-sensitive learning and oversampling to reduce data imbalance. Journal of medical Internet research, 21(3):e11990, 2019.

[2] R. Hu, J. Chen, and L. Zhou. A transformer-based deep neural network for arrhythmia detection using continuous ecg signals. Computers in Biology and Medicine, 144:105325, 2022.

[3] P. Shimpi, S. Shah, M. Shroff, and A. Godbole. A machine learning approach for the classification of cardiac arrhythmia. In 2017 international conference on computing methodologies and communication (ICCMC), pages 603–607. IEEE, 2017.

[4] B. Wang, C. Liu, C. Hu, X. Liu, and J. Cao. Arrhythmia classification with heartbeat-aware transformer. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1025–1029. IEEE, 2021.

[5] G. Yan, S. Liang, Y. Zhang, and F. Liu. Fusing transformer model with temporal features for ecg heartbeat classification. In 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 898–905. IEEE, 2019.