A Work Project, presented as part of the requirements for the Award of a Master's degree in Management from the Nova School of Business and Economics.

MACHINE LEARNING AND DEEP LEARNING IN HEALTHCARE: ADVANCING CARDIAC ARRHYTHMIA CLASSIFICATION IN HEALTHCARE ANALYTICS

ALEXANDER MEHLER

Work project carried out under the supervision of:

Rodrigo Belo

17/01/2024

Abstract (100 words maximum)

Cardiac arrhythmias, a global leading disease cause, necessitate rapid, efficient diagnosis. Shifting from traditional manual electrocardiogram analysis to machine learning approaches offers enhanced efficiency and accuracy in detection. However, literature research has shown that long training times and a lack of practical suitability have made implementation difficult to date. Three prototypes were developed and tested; the results were then used to optimize the most promising model further. The optimized CNN achieved an overall classification accuracy of 98.53%. The results are tested for their applicability in a practical context, evaluated, and compared against existing approaches, resulting in above-average classification outcomes.

Keywords:

Predictive Modeling; Convolutional Neural Networks; Deep Learning; Arrhythmia; Classification; Hybrid Models

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

1. Introduction

Every year, cardiovascular diseases cause almost 20 million deaths worldwide (O'Riordan 2022). According to a study by the American Heart Association, this figure is expected to increase by 23.56% by 2030, bringing the annual death toll to 23.6 million per year (Angell et al. 2020). Making this the leading cause of death worldwide shows that cardiovascular diseases are omnipresent in our daily life, known and experienced in a family context or from personal experience (O'Riordan 2022). One form of these cardiovascular diseases is arrhythmia. Although this type of disease is well known, one may have difficulty explaining what it implies. So, the question arises: what is arrhythmia, and how can it be treated effectively? According to the medical dictionary, "Arrhythmia" refers to a group of conditions reflecting electrical impulses that vary from the typical sequence, causing the heart to beat erratically (Cambridge Heart Clinic 2019). Arrhythmia occurs in all age groups (Khanal et al. 2023). Therefore, early detection and treatment are essential to survive arrhythmia, mainly because symptoms often occur unnoticed initially. The diagnosis is made using ECG data, which, because of their noninvasive nature, acts as a convenient diagnostic tool.

Due to its increasing prevalence, research on the detection and classification of cardiovascular diseases has gained more interest over the past decades, especially in machine learning (Chen et al. 2022). The emergence of computational resources and the development of intelligent devices, achieving continuous and remote monitoring of ECG, allows researchers to believe that diagnostic systems based on machine learning can minimize the burden of instinctive uncertainty of experts, potentially leading to a misdiagnosis (Javaid et al. 2022). Therefore, a shift towards a less time-consuming and laborious option is necessary (Appendix 1). A study by Sturman et al. (2020) shows that classification using deep learning can generate higher accuracy and efficiency than a cardiologist's classification.

This study focuses on machine and deep learning models and their underlying mechanisms: the development of a vigorous method that can be used in practice. It is well known that modern machine learning approaches have the potential to recognize precise patterns in ECG data due to their robustness and generalizability. According to comparative studies, such as that by Madani et al. (2018), there are yet few models that are both practicable and precise in their predictive ability. Hence, this study's goal is to contribute to closing the research gap by developing a model using minimal preprocessing and receiving over-benchmark results. Therefore, the present study concerns the research question, "How effectively can machine learning methods be used to detect and classify different types of cardiac arrhythmias based on ECG data?".

This thesis is organized as follows. The *Literature Review* states the current status quo of researcher findings. The experimental setup consists of data set, preprocessing, prototyping, and fine-tuning, described in the *Methodology* section. The following results, interpretation, and the corresponding viability analysis and implications for science and practice are presented in the *Analysis and Discussion* section. Finally, the *Conclusion and Prospects for Future Work* section presents a summary and concrete approaches for future work.

2. Literature Review

In recent years, many well-designed methods have been proposed for arrhythmia classification. The existing models can be roughly divided into traditional machine learning approaches, such as Random Forest or Logistic Regression, or more modern approaches considering deep learning. Differentiation proves to be helpful, as both traditional and modern approaches have different degrees of complexity, data requirements, and interpretability. For instance, lower degrees of complexity in machine learning models lead to improved interpretability but, at the same time, reduced performance. Therefore, a differentiation was made in order to enable a

more effective comparison between developed and optimized machine and deep learning models in the further course of the study. This chapter presents previous research efforts to classify cardiac arrhythmia. "Accuracy" is used to assess the algorithm's performance and benchmark it against existing evidence, even though additional metrics are used to assess custom-built models (cf. Chapter 4).

Traditional machine learning approaches offer the advantage that they often require less computing time and are highly relevant in time-limited contexts than deep learning models (Taye 2023). This advantage was also considered in the work of Gupta et al. (2021), in which they achieved a generalization accuracy of 77.40% by applying contemporary literature such as Naïve Bayes, Support Vector Machine, Random Forest, etc.. The experimental results make it clear that the application of individual traditional algorithms is not sufficient for precise forecasts. For illustration, recent machine learning models, as presented in the study by Barboza, Kimura and Altman (2017), demonstrate around 10% higher predictive accuracy of different use cases than traditional methods. Yet, combining Support Vector Machine and Random Forest leads to an improved generalization accuracy. This increase is driven by the complementary strengths of both methods, which make it feasible to capture different patterns in the data. Contrary to the approach prevalent in most existing literature, the developers have categorized errors into 16 distinct classifications. Confusion matrices show that both models struggle to differentiate normal heartbeats from supraventricular arrhythmia. Luz et al. (2013) used an Optimal Path Forest classifier and achieved a detection rate of 90.10% for arrhythmias. Although this detection rate is lower than the performance of some other models, the classifier is characterized by its time efficiency, as it can perform calculations significantly faster than existing models (Papa et al. 2012). Another study investigates an optimized combination of feature selection and Random Forest classifier for automatic heartbeat classification systems, especially for resource-limited applications (Saenz-Cogollo and Agelli 2020). Features were

selected using a filtering method based on the mutual information ranking criterion, where normalized R-R intervals and features related to QRS complex width were identified as the most discriminative (ibid.). The method achieved an overall accuracy of 96.14% on the MIT-BIH Arrhythmia Database. These results show that conventional methods can deliver promising results when applied in tailored contexts.

Although some researchers rely on modeling using traditional approaches, others point out that modern deep-learning models have more praising prospects. Haasan et al. (2014) obtained significant results by implementing a fuzzy c-means based clustering method (FCM) within a probabilistic neural network. The average classification accuracy obtained was 97.54% when applying an FCM-clustered multilayer neural network. These results were made possible by extracting and reducing individual ECG data using clustering, which served as input to neural network classifiers. Researchers Nasim and Kim adopted a similar methodology, utilizing an evolutionary optimization technique, specifically differential evolution, for feature classification and employing a probabilistic neural network in their study (Nasim and Kim 2022). Using this method, they could dismiss the computationally intensive algorithms frequently found in the literature. Through direct reduction and optimization of the features prior to their introduction into the network, they achieved an overall classification accuracy of 99.33%. Irfan et al. (2022) integrate different neural networks by individually stacking similar layers in each network. This has shown that two existing problems, high training time and manual feature selection, could be significantly reduced, leading to an accuracy of 98.41%. Parvaneh et al. conducted a comparative analysis of various traditional deep learning algorithms, focusing on the impact of neural network depth on performance. Their research indicates that convolutional neural networks (CNN) are promising in addressing classification challenges.

Comparative studies, among others, by Madani et al. (2018), in which different models were compared show that traditional classifiers such as Random Forest or Support Vector Machine have significantly lower accuracy than models based on deep learning approaches such as Convolutional Neural Networks. These studies underscore the importance of further research in modern modeling approaches and highlight performance in terms of computational effort and accuracy compatibility compared to traditional classifiers. More recently, the authors have insistently partnered with deep learning models for scalable, robust, and efficient arrhythmia heartbeat classification (Essa and Xie 2021). The designed models strived to optimize automatic feature extraction, minimize overall computational overhead, and provide increased accuracy and precision. For instance, Sarfraz et al. (2015) enhanced ECG pattern recognition performance by incorporating Independent Component Analysis in conjunction with R-R interval and QRS segment power as inputs to a neural network. Their results demonstrate improved recognition accuracy with a reduction in data preparation time.

The studies mentioned above prove the existence of sophisticated classification models, which, however, need to be critically evaluated regarding their robustness and effectiveness in real-time applications. Appendix 2 summarizes the published models on the MIT-BIH data set developed by different researchers and serves as a benchmark overview.

3. Methodology

The quantitative research methodology in this study focused on analyzing numerical ECG data to identify abnormal patterns and answer the research question concerning whether automated arrhythmia detection can be optimized. The cross-industry standard process for data mining is employed to ensure that industry-wide standards are adhered to benchmark with existing studies (Rivo et al. 2012). The process up to the final model development consists of three steps: first, the data inspection and subsequent data preparation. Second, prototyping based on the machine

and deep learning approaches. Followed by the third step, the optimization of the chosen model to finetune parameters, and model architecture to generate robust and reliable outputs.

3.1 Data Collection and Preprocessing

This work was based on the publicly available MIT-BIH Arrhythmia data set. This data set comprises 48 half-hour sections of two-channel ECG recordings collected from 47 individuals at Boston Beth Israel Hospital between 1975 and 1979 (Moody and Mark 2001). The dataset structure is divided into two parts: The first consists of 23 randomly selected data points from 4,000 ECG recordings, with 60% generated from inpatient data and 40% from outpatient data. The second part comprises 25 recordings from the same collection, representing less frequent but clinically significant arrhythmias. The bipartition of the data set was done to increase the representativeness of the data set, as these specific arrhythmias would not be adequately represented in a small random sample. Furthermore, the data set contains annotations that divide the ECG data into classes to facilitate the evaluation of arrhythmias (Moody and Mark 2001).

ECG data was processed and analyzed using the Waveform Database Software Package (WFDB). This toolset allowed for the precise extraction of annotations that contained critical information about the timing and clinical classification of each heartbeat (Goldberg et al. 2000). The selection of the WFDB package for the data transformation step was based on its seamless integration with the PhysioNet platform, its wide recognition within the biomedical research community, and its extensive documentation that facilitates implementation (Xie et al. 2023). With a total of 109,446 data points, the data set provides adequate data for training traditional and modern models like deep neural networks. The data were classified according to the AAMI / ANSI standards to ensure high comparability with existing studies (American National Standards Institute and Association for the Advancement of Medical Instrumentation 1998). An overview of the instances from the different classes can be found below.

TABLE 1 – ECG classification according to AAAMI / ANSI standards

Classes	Beat type	Mapping to MIT-BIH	Number of instances
N	Normal beat	e, R, N, L, j	90,631
S	Supraventricular Ectopic-Beat	J, A, a, S	2,781
V	Ventricular Ectopic-Beat	V, E	7,239
F	Fusion-Beat	F	803
Q	Unknown Beat	/, Q, f	8,043

As expected in anomaly detection, the data set has a highly uneven distribution, as shown in Appendix 3. To counteract dominance in the prediction of the predominant class, a weighting factor was implemented to scale the loss function of the training and validation dataset, which is based on the calculation, namely

$$Class\ weight = \frac{n_{samples}}{n_{classes} \times np.bincount(y)}, \qquad (1)$$

used in Scikit-learn (Pedregosa 2011). No further steps were carried out for prototyping to see how the respective models perform without first tuning parameters or making fine adjustments to the data set to align with the stated research aims.

3.2 Development of Exploratory Models

The models were developed using the back-end Python libraries from NumPy (Harris et al. 2020), Pytorch (Paszke et al. 2019), Seaborn (Waskom 2021), Pandas (Pandas development team 2020), Scikit-learn (Pedregosa 2011) and Matplotlib (Hunter 2007). In addition, deep learning libraries, particularly Keras and TensorFlow, were used for the implementation as they allow an accessible installation on different operating systems and provide a solid introduction to neural networks (Keras-Team 2015) (Abadi et al. 2015). Based on the pre-processing methods described in Chapter 3.1, the same split in terms of training and test set (80:20) was chosen for all models. Modeling was done using pipelines, which were selected mainly due to

their better reproducibility, workflow simplification, and easier transferability to production, which is one of the critical elements evaluated in the viability analysis (Hapke and Nelson 2020).

3.2.1 Prototype Model 1 – Stacking (RandomForest, XGBoost, Logistic Regression)

The first model tested was a stacking model. This ensemble learning approach computes an average prediction obtained from several base models, each followed by a single final prediction that shows more predictive power than individual predictions (Srinivasan 2022). In the base model of this work, a Random Forest classifier and an XGBoost classifier were used in combination with a Logistic Regression meta-learner, which attempts to make the final classification based on the input from the base models.

Within the stacking model, Random Forest acts as a base model to recognize complex, non-linear relationships in the data (Breiman 2001). This classifier was configured by the parameter *n_estimators* so that 100 trees are created to generate better generalizability and stability of the predictions. When splitting each node during the construction of a tree, the best split is found either from all the features entered or from a random subset - during development, the default value, i.e., the square root of the number of features, was used. Random forest was also chosen due to its ability to construct decision trees in parallel. For a faster calculation, *n_jobs = -1* was used to perform the calculations on all CPU cores. For better reproducibility of the results, *random_state* was set to 42. The parameters of the model were selected based on the findings of the literature review, which showed that these parameters were often used in isolation and are now combined here.

A XGBoost classifier was used as a further base model. This gradient-boosting approach is often used for classification tasks, as it offers a more accurate approach than the Random Forest. Especially for inbalanced datasets, as it is the case for arrhythmia detection, the XGBoost shows

its strength in assigning a higher preference and weighting to the anomaly in the following iterations and is thus able to predict classes with lower probability of occurrence as accurately as classes with a high probability of occurrence (Moore and Bell 2022). Since a division into five different error classes is made, i.e., a multiclass division, *num_class* = 5 was set. In addition, the best parameter configuration was searched by using GridSearch for the parameters *max_depth*, as determination of the depth of the respective trees, *learning_rate*, as step size for updating the weights, and *n_estimators*, size for the number of trees.

Each base model undergoes independent training on the training data, followed by the training of the stacking model. During training, the Random Forest classifier creates 100 decision trees based on different data set samples, whereas the XGBoost classifier creates trees sequentially. Each subsequent tree attempts to correct the errors of the previous one. After training, the two base models are used to make predictions on the same training data set. The final estimator, in this case, Logistic Regression, attempts to use these predictions from Random Forest and XGBoost outputs as input parameters to create a model that follows the principle of second-order learning.

3.2.2 Prototype Model 2 – Light Gradient Boosting Machine (LightGBM)

The second model implemented uses a Light Gradient Boosting Machine (LightGBM) approach. This specific methodology is selected based on the recognized efficiency of LightGBM in processing large datasets. In addition, LightGBM is characterized by its ability to deal effectively with unevenly distributed data, as well as with the anomalies at hand. LightGBM is an advanced gradient-boosting framework that differs from the conventional sorting-based decision tree learning algorithm using a highly optimized histogram-based algorithm (Microsoft Research 2023). The calculation method employed is the one-sided

gradient-based sampling method, which can be reflected in equation 2, f representing the model's output, y the true label and w the weight for the selected instance.

$$g = w\,(f - y); h = w \qquad (2)$$

The configured LightGBM classifier includes parameters such as *num_leaves*, the maximum number of leaves per base learner, *learning_rate*, step size at which the algorithm makes updates to the model weights, *n_estimators*, and the number of boosted trees to be trained. In this model, similar to the previously discussed stacking model, GridSearch is utilized to determine the optimal parameter configurations. Attention was also paid to the model's input variables, as LightGBM processes the data in a one-dimensional, label-coded form (Ke et al. 2017). The model is trained with optimal parameters identified through GridSearch to ensure optimal performance during classification.

3.2.3 Prototype Model 3 – Convolutional Neural Network

A convolutional neural network was developed as the third model, representing a promising approach to processing time series data.

The first layer of the network structure is a one-dimensional convolutional layer (Conv1D) with 64 filters and a kernel size of six. This configuration allows the network to extract significant features from the input data. Using Rectified Linear Unit (ReLU) as an activation function, this layer introduces an effective non-linearity in learning. This makes the model better at recognizing complexities, as well as patterns in the data and deriving more sophisticated results based on them. This is enhanced by adding a batch normalization layer that adapts mean and variance change with training time and data standardization. The training is accelerated by normalizing the output of the previous layer. A Max Pooling layer with a pool size of three then

reduces the spatial size of the output. This helps to reduce the number of features processed and creates a larger optimization window for the CNN. Subsequently, using a flatten layer, the feature map obtained by Max Pooling can be converted into a one-dimensional array so that the Dense layer can process the data. The dense layer of 64 neurons is provided with a ReLU activation function, which ensures the further processing and combining of the extracted features. A dropout layer with a rate of 0.4 was added to reduce the risk of overfitting. During training, 40% of the total number of neurons are randomly switched off. This measure promotes the stability of the model in the event of minor changes and disturbances within the input data. As delineated in the literature review, this specific rate is frequently employed as a dropout rate. Consequently, it has been adopted for this model to facilitate benchmark comparisons. The network's last layer has another dense layer with five neurons - corresponding to the number of arrhythmia classes. Using the SoftMax activation function, it is possible to calculate probabilities for each class and thus interpret the model output probabilistically.

3.3 Model Optimization

The selection of the convolutional neural network for the optimization is explained in sufficient depth in the *Analysis and Discussion* section. The optimized model developed for this thesis comprises three convolutional layers, followed by Rectified Linear Unit activation, batch normalization, dropout layer, and Max Pooling. Furthermore, a long short-term memory (LSTM) was tested to feed the data extracted from the CNN into an LSTM for classification. The implementation did not show significant improvements in the model, which is why its use was not pursued further - the results of the empirical observation can be found in Appendix 4. Each of the three existing convolutional blocks consists of four layers, which are followed by a dense and output layer. In alignment with the prototype model, a Conv1D layer was selected, which is characterized by a convolution kernel that integrates with the layer input along a single

dimension and thus offers advantages in the processing of time series data. Parameters of the Conv1D layer were set to *padding* = *same*, for a uniform input output shape, along with the activation function ReLU. The activation function can map the resulting values of a layer between a specific range of values. According to Lapid and Sipper (2022), ReLU functions are convolutional neural networks' most frequently used activation functions. The formula below shows that only the positive values pass, while the negative values are transformed to 0.

$$f(x) = \max(0, x) \qquad (3)$$

The use of ReLU is particularly useful because it facilitates identifying anomalies; other activation functions such as Elu or Sigmoid were not pursued further due to tested lower performance. Although, activation functions like Elu showed in theorical settings better results. Each of the three blocks comprises a progressively doubling number of kernels, starting with 32 in the first layer, 64 in the second, and culminating at 128 in the third layer, preserving features and hierarchical relationships. After each Conv1D layer, a batch normalization is utilized to reduce the internal covariance shifts and thus to standardize the data adaptively. Compared to the prototype, the optimized model uses a dropout rate of 10% to prevent overfitting and increase robustness to small variations in the input data. This layer is followed by a Max Pooling layer with the parameters 2x2 and strides two to down sample the features, which increases the computational efficiency.

The data then passes through a flattening layer and is converted from the multidimensional feature map into a one-dimensional vector. This operation is followed by a non-linear activation function, significantly improving the network's ability to learn and abstract complex patterns. The SoftMax function converts the output of the neural network into a probability distribution over the defined error classes. The SoftMax output for a given input is

$$s\,(x_i) = \frac{e^{x_i}}{\Sigma_{j=1}^{n} e^{x_j}} \qquad (5)$$

that ensures that the sum of all probabilities for the different classes equals one. In application, this function provides a precise probability estimate for each class of arrhythmias.

Finally, the model is complied with the optimization algorithm Adam, which uses adaptive learning rates for different parameters. Through testing, an optimal learning rate of 0.001 could be determined, which allows convergent steps towards the global minimum of the loss function without oscillating or stagnating excessively. Furthermore, decay was set to 1e-6 to slowly reduce the learning rate over time. Both settings are intended to help detect subtle patterns within the data. Early stopping is used in the training process to prevent missing improvement on the validation set.

4. Analysis and Discussion

The results were evaluated using the following performance metrics: Accuracy, Precision, Recall, and F1-Score. Additionally, the Matthews Correlation Coefficient was utilized as it is considered a more reliable statistical measure that produces a high score only if the prediction obtained good results in all four confusion matrix categories (Bordoloi and Biswas 2023). To avoid the occurrence of the accuracy paradox, a holistic approach was chosen, involving multiple metrics for evaluation rather than solely focusing on one value, accuracy. These formulas of the performance metrics can be found in Appendix 5. Particular attention is paid to minimizing type II errors (false-negative results), especially in the classification of ventricular (V) and supraventricular ectopic (S) beats, as these can indicate serious cardiac abnormalities. Overlooking these arrhythmias can lead to delayed or missed treatment and, therefore, increased health risks. In contrast, although type I errors (false-positive results) should be avoided to reduce unnecessary medical interventions and patient distress, they do not pose a risk to the

patient due to their non-invasive nature. Careful monitoring is required in the category of unknown beats (Q), as these unclear findings could indicate previously unrecognized types of arrhythmias. Therefore, the optimization of the classification model should aim at a balanced reduction of both types of errors, with a particular focus on avoiding type II errors in critical categories.

4.1 Evaluation of Prototype Models

In the previous chapter, we delineated the methodology employed to transform our pre-processed dataset by applying multiple classification models. These models have unique settings that affect the respective classifiers performance. The purpose of this prototype modeling was to make some preliminary inferences on potential algorithms and respective performance indicators. The evaluation section noted that various measures were taken into account to come up with a clear image of the strengths and the weaknesses of the models discussed earlier. However, an emphasis of the assessment is placed on providing answers to the research question that underlies a practical algorithm to be modeled, thereby allowing for automatic detection of the arrhythmias. The computational costs, as well as the traceability of the results, are also discussed to make them applicable.

TABLE 2 - Results Prototype Stacking Model

Classes	Accuracy	Precision	Recall	F1	MCC
N	0.9727	0.9685	0.9995	0.9838	0.9025
S	0.9883	0.9870	0.5450	0.7022	0.7289
V	0.9893	0.9919	0.8453	0.9128	0.9104
F	0.9966	0.8537	0.6481	0.7368	0.7422
Q	0.9961	0.9974	0.9490	0.9726	0.9708

The stacking model shows a good detection of regular beats (N) with an accuracy of 97.27% and an MCC of 90.25%, indicating a reliable identification of the most common heartbeat type. For supraventricular ectopic beats (S), despite the high accuracy of 98.83%, the model achieves only an MCC of 72.89%, which underlines the need to improve the distinctiveness in this class. The Confusion Matrix, which can be found in Appendix 6, shows that mistakes are mainly made with the category of normal heartbeats (cf. chapter 4, type II error). One reason for this may be the low level of data preparation, as the ECG curve for category S errors is very similar to that of normal heartbeats (Appendix 7). The detection of ventricular ectopic beats (V) is reliable, with an accuracy of 98.93% and an MCC of 91.04%, even though there is room for increased Recall. Whilst the model detects fusion beats (F) with an accuracy of 99.66%, the MCC of 74.22% indicates difficulties in precise classification; possible reasons for this have already been explained above. Overall, the stacking model solidly performs across all classes, focusing on improving precision and reducing misclassifications, especially in classes S and F.

TABLE 3 - Results Prototype LightGBM

Classes	Accuracy	Precision	Recall	F1	MCC
N	0.8876	0.9871	0.8756	0.9280	0.7017
S	0.9514	0.3198	0.8094	0.4585	0.4905
V	0.9671	0.6930	0.9012	0.7835	0.7738
F	0.9733	0.1994	0.8642	0.3241	0.4078
Q	0.9841	0.8447	0.9608	0.8990	0.8927

The LightGBM model features a solid accuracy of 88.76% and an MCC of 70.17% in detecting normal beats, indicating acceptable reliability. For supraventricular ectopic beat, the model achieves a higher accuracy of 95.14%, but the low MCC of 49.05% signals large room for improvement in precision. The detection of ventricular ectopic beats with an MCC of 77.38% and an accuracy of 96.71% is promising, making the model in most cases reliable for critical

heart rate classifications. Despite the accuracy of 97.33% in class F, the MCC of 40.78% shows significant limitations in its distinctiveness, as it is similarly seen in the stacking model - now, however, significantly performing lower. The Q category is handled above average, with an MCC of 89.27% and an accuracy of 98.41%, which underlines the strength of the LightGBM model in assigning undefined beats. The reason for this could be the gradient boosting approach which makes it possible to segment ECG data in a fine-grained way and thus recognize more subtle patterns which, for example, the stacking model could not recognize. Despite the nuanced nature of electrocardiographic data, where subtle patterns or anomalies can easily be overlooked, the use of LightGBM for pattern recognition may, in some cases, result in overinterpretation of the data.

TABLE 4 - Results Prototype Convolutional Neural Network

Classes	Accuracy	Precision	Recall	F1	MCC
N	0.9783	0.9880	0.9857	0.9868	0.9242
S	0.9881	0.7701	0.7590	0.7645	0.7584
V	0.9921	0.9327	0.9482	0.9404	0.9362
F	0.9959	0.6915	0.8025	0.7429	0.7429
Q	0.9973	0.9813	0.9813	0.9813	0.9799

The CNN model strongly detects regular beats with an accuracy of 97.83% and a precision of 98.80%, indicating an effective filtering of false positives, confirmed by an MCC of 92.42%. On the other hand, class S has a precision of 77.01% and an MCC of 75.84%, indicating a tendency towards false positive results. Ventricular ectopic beats are efficiently identified with an accuracy of 99.21% and a high MCC of 93.62%, denoting reliable predictive power in critical categories. The class F (fusion beat), with an MCC of 74.29% despite a high accuracy of 99.59%, shows strong weaknesses in precision, which suggests a risk of confusion with other beat types - the primary misclassifications occur with the categories N and V (Appendix 7).

The strong performance for unkown beats with an MCC of 97.99% shows the strength of the model in identifying unclassified beats - which is of lesser importance, however, as models are intended to be used primarily for the detection of arrhythmias and not to imply to the patient that it is an unclassified beat.

In conclusion, the convolutional neural network shows higher MCC values than the other models across all groups, indicating superior performance in the balance between sensitivity and specificity. Especially in the present context, where avoiding misdiagnoses, especially those that incorrectly classify arrhythmias as normal heartbeats, is a top priority, the MCC is a good indicator. Several well-founded rational considerations can underpin the preference for a convolutional neural network in this context. Looking first at the LightGBM model, it can be stated that it has already been subjected to a tuning process, albeit with a limited variety of parameters. This suggests that the LightGBM model is close to its maximum performance. Further fine-tuning could, therefore, only lead to minor increases in performance. In contrast, the short computation time of the model should be emphasized, which is highly relevant for practical implementation. In contrast, CNNs are characterized by their pronounced adaptability and efficiency in processing signals. Given that the current CNN model is still in its basic form, this suggests significant potential for improvement through further tuning. Especially in the effective detection of complex features and generalizability to different data sets, CNNs could be superior to the Stacking or LGBM model. Looking again at the results of the stacking model, it can be stated that type II errors, in particular, occurred to a significant extent. This observation, which was already discussed in depth in the fourth chapter, considerably undermines the practical applicability of the model. Moreover, the adaptability of the CNN architecture is designed for use with different datasets and formats. It means that this adjustment ability is especially important to apply the model in different cases. The use of CNNs is also a dynamic area of research that allows researchers to incorporate the newest scientific data and

methods into the model's ongoing work. These considerations add to the justification of pursuing the CNN model in detail and fine-tuning it.

4.2 Performance Assessments of the optimized CNN

As described in Chapter 3.3, the optimized model is based on the basic CNN model to improve classification performance. The results show differentiated performances across all classes, with strong performance in classes N, V and Q being particularly noteworthy. Similarly, the MMC value is consistently above average, which indicates a strong and balanced classification performance - here likewise, predominantly for categories Q and N. The following table provides an overview of the exact results.

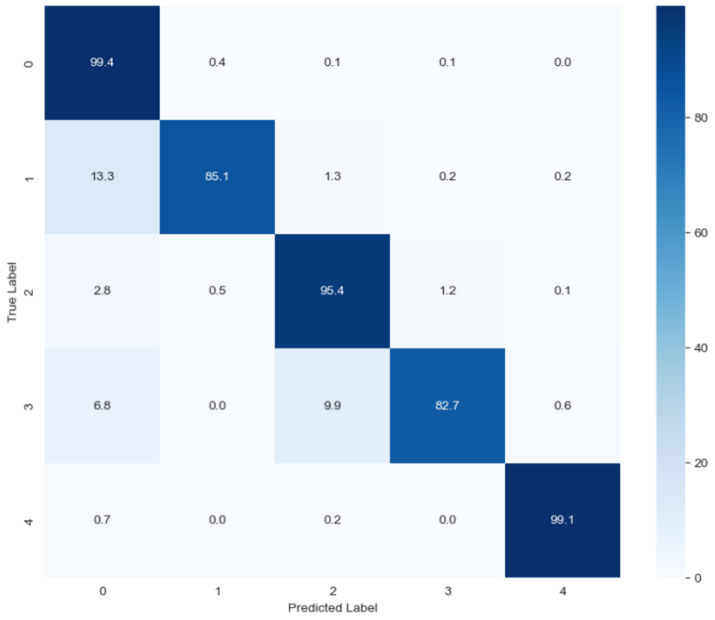
TABLE 5 - Results of Optimized Convolutional Neural Network

Classes	Accuracy	Precision	Recall	F1	MCC
N	0.9888	0.9925	0.9940	0.9932	0.9607
S	0.9929	0.8663	0.8507	0.8584	0.8548
V	0.9951	0.9705	0.9544	0.9624	0.9598
F	0.9971	0.7929	0.8272	0.8097	0.8084
Q	0.9987	0.9919	0.9907	0.9913	0.9906

The model shows a consistently above-average performance for the normal ECG pattern category. All of the available values are above 98%, except for the Matthews Correlation Coefficient with a score of 96.07%. These superior scores could be partly explained by the fact that the clear and consistent features of normal ECG patterns simplify the classification task, making classification easier. The hypothesis that the improved modeling of the structures results from an overrepresentation of normal ECG data is rejected due to the implementation of weighting techniques presented previously in Chapter 3.1.

A differentiated view is required if a deeper focus is placed on areas in which the model shows lower values for precision and recall. The results achieved in classes S and F are particularly noticeable here, which, when looking at the ECG curves in Appendix 6, may indicate that the intrinsic variability within these arrhythmia types makes classification rather difficult. This could be attributed to the heterogeneity of arrhythmia patterns within a class, which poses additional challenges for the model. A confusion matrix can be used to see which categories are frequently confused among each other. The matrix corresponding to the optimized model can be found below.

FIGURE 1: Confusion Matrix CNN (optimized)



For the two remaining classes, i.e. classes V and Q, superior values can be achieved consistently. Once again, Appendix 7 shows that the shape of the curves differs markedly from that of a normal heartbeat. The appearance of these distinctive features suggests that the model is capable of capturing more subtle distinguishing features, indicating an appropriate configuration of filters and neurons in the convolutional and dense layers. Similarly, it also provides a potential explanation for why the model has more difficulty distinguishing these two

types from each other than those with different curves. The detailed analysis of the confusion matrix reveals similar results to those already explained. Two class results should be emphasized here: the supraventricular ectopic beats and the fusion beats. In both cases, high confusion rates with class N can be observed - a particularly undesirable scenario in the clinical application context (cf. Chapter 4). Although any form of misclassification is suboptimal, confusion with other types is considered less problematic. This is due to the fact that people who are actually ill are misclassified as healthy and may, therefore, not be treated. However, it should be highlighted that despite the high misclassification rate of class S, the precision metric is 86.63%, which indicates average performance.

FIGURE 2: Training/ Validation Loss & Accuracy of CNN (optimized)

The significant reduction in training loss and the simultaneous increase in training accuracy indicate an efficient learning performance of the network, potentially favored by the use of Adam optimization (learning rate: 0.001, batch size: 128). In contrast, an increasing validation loss and an inconsistent validation accuracy signal the phenomenon of overfitting. Hypothetically, an overly complex network architecture and suboptimal dropout rates (currently at 0.1) could serve as causal factors. Various settings were systematically tested to

determine the optimum configuration of the dropout rate. The empirical results, presented in detail in Appendix 8, show that a dropout rate of 10% delivered the most effective results. Furthermore, increasing the early-stopping patience to 10 epochs was considered to allow the network to adapt to performance fluctuations and promote optimal results. These modified optimization strategies resulted in superior results, as limitations in performance or applicability would have been observed otherwise (Kıranyaz, İnce, and Gabbouj 2016). A final comparison of the results with those of the initial CNN model shows that a significant improvement was recorded across all error classes and metrics, manifesting itself in a more precise classification of the arrhythmia categories.

4.3 Viability and Implications of Optimized CNN

Results are only relevant if they are viable and can be deployed in the problem context. If this is not the case, the model may show statistically promising data, but these will not find any actual application. This is particularly important in the context under consideration, as any wrong decision could have far-reaching consequences - up to and including the death of a sick, untreated patient. Therefore, viability and implications are being considered in the following.

The implementation of CNNs in healthcare, especially in the diagnosis of cardiac arrhythmias, promises significant benefits, although it must be thoroughly analyzed, taking into account costs, risks, and data protection aspects. On the one hand, using CNNs in ECG data analysis enables time savings of up to 50%, reducing the average diagnosis time per ECG from 60 to 30 minutes (Barth & Barth 2023). This increase in efficiency can improve diagnostic capabilities, particularly in peripheral regions, and provide higher diagnostic accuracy of 98.53% compared to the previous 86% of human decision-making (Newman-Toker et al. 2020a).

However, the cost of implementing such systems is not negligible. Innovative approaches, such as using the residual class system in hardware implementation, can reduce hardware costs by

up to 37.78% (Valueva et al. 2020). These approaches help to reduce the initial and ongoing costs of applying CNNs, which is crucial for sustainable implementation in healthcare systems. In addition, a 2016 study found that once trained, dedicated CNNs can quickly and accurately classify ECG recordings for individual patients - which can also lead to resource savings.

In addition to costs, the risks and challenges associated with the accuracy and reliability of CNNs must also be considered. For example, the likelihood of system failures and misdiagnosis, despite low rates, cannot be ignored and requires constant monitoring and improvement of the systems. According to Amazon Web Services, the probability of a system failure is currently merely 1% (AWS 2023). Another important aspect is compliance with the General Data Protection Regulation (GDPR) in the EU, which poses particular challenges for processing sensitive healthcare data. Compliance with the GDPR requires significant efforts in terms of data security and protection, which demands additional costs and resources. Kammüller (2018) highlights that violations of the GDPR can result in fines of up to 20 million euros, which drives users to ensure proof of compliance through formal modeling and analysis - while at the same time entailing a high-risk potential.

The far-reaching implications of this technology for clinical practice, public health and medical education are nevertheless significant. Through early detection of cardiac risk patients, preventive measures can be taken that could improve the overall health of the population and reduce the risk of serious cardiac events. In addition, models such as the one developed could play an essential role in the training of healthcare professionals by facilitating the interpretation of ECG data and improving diagnostic skills. Referring to the beginning of the thesis, this means fewer deaths related to cardiovascular disease due to early detection.

5. Conclusion and Prospects for Future Work

This chapter summarizes the research, draws conclusions from that, and indicates some of the implications of the findings. Limitations of the study, as well as suggestions for further research in this context, are considered.

This study investigated and evaluated different ensemble and deep learning techniques to accurately classify arrhythmia. The publicly available data was first preprocessed accordingly to be divided into training and test sets. Three prototype models were initially developed to provide preliminary results for selecting the most promising one. The convolutional neural network showed the most promising results of the developed models. However, the overall accuracy of the other models, the Stacking Model and the LightGBM was generally acceptable for pursuing. The further developed CNN was optimized to such an extent that it was able to achieve an overall accuracy of 98.53%, MMC of 91.49%, and increase the classification within the individual arrhythmia classes to 99.4%, 85.1%, 95.4%, 82.7%, and 99.1% respectively. However, it was also found that using residual blocks positively affects the flow of information through the network by coping with the vanishing gradient in the deep network but also harbors the risk of overfitting.

Furthermore, the data set comprises approximately 110,000 data points, which, due to the size of the data set and the additional human annotations for the arrhythmia classifications, can make it challenging to train a model that works consistently and reliably across different data types. However, the research question “How effectively can machine learning methods be used to detect and classify different types of cardiac arrhythmias based on ECG data?” can be answered by this study’s results. According to the results, the proposed model shows better results than some other studies in ECG classification. It may serve as a potential tool for aiding ECG detection and classification. Building on the findings of this work, future research could be carried out focusing on real-time recording, receiving the ECG signal, and processing it

simultaneously to get closer to this goal. This would offer the advantage that by expanding the ECG signal database, the model can better recognize patterns and thus improve classification performance. The simultaneous use of an automated adjustment of the model based on feedback, e.g. doctors validating or rejecting the decision, could also be used to improve models performance. Implementing such an approach would strengthen the potential for large-scale use within the population and thus actively contribute to reducing cardiovascular disease through early identification. An additional approach to optimizing the developed convolutional neural network involves fine-tuning with an increased number of parameters and increasing the complexity of the network structure. However, this requires significantly higher computing capacities. During the development phase of the model, such methods were considered, but they had to be adapted due to limited computing resources. More powerful computers offer extended computing capacities that can be used effectively for such optimizations. For example, a less limited number of parameters can be tested in order to obtain more precise configurations of the models. Furthermore, the combination with other diagnostic data such as blood tests or patient anamnesis also represents an opportunity to better visualize correlations. However, it is important to ensure that future developments comply with ethical guidelines and data protection regulations, especially with regard to sensitive health data.

A comprehensive literature review has shown that this work provides important insights into current research. To establish a direct link to the results of this work, it is recommended to use the developed neural network as a starting point for further operations. It serves to reduce existing classification difficulties, and thus contribute to increased performance. Due to the selection of a CNN, it is also possible to adapt the structure of the network with minimal coding effort so that the latest findings or parameters can be incorporated - it thus serves as a toolbox that can be continuously expanded and adapted.

References

- Abadi, MartíN, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, et al. 2015. “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.” arXiv. Ithaca, NY: Cornell University.

- Alfaras, M.; Soriano, M.C.; Ortín, S. 2019. “A fast machine learning model for ECG-based heartbeat classification and arrhythmia detection.” Frontiers in Physics 7: 103. Lausanne: Frontiers Media S.A..

- American National Standards Institute and Association for the Advancement of Medical Instrumentation. 1998. “ANSI/AAMI EC38:1998, Ambulatory Electrocardiographs.” Arlington, VA: Association for the Advancement of Medical Instrumentation.

- Angell, Sonia Y., Michael V. McConnell, Cheryl A.M. Anderson, Kirsten Bibbins-Domingo, Douglas S. Boyle, Simon Capewell, Majid Ezzati, et al. 2020. “The American Heart Association 2030 Impact Goal: A Presidential Advisory from the American Heart Association.” Circulation 141 (9).

- AWS. 2023. “Availability with Redundancy - Availability and Beyond: Understanding and Improving the Resilience of Distributed Systems on AWS.” Seattle, WA: Amazon Web Services.

- Barboza, Flavio, H. Kimura, und E. Altman. 2017. "Machine learning models and bankruptcy prediction." Expert Systems with Applications 83: 405–417. Amsterdam: Elsevier Ltd..

- Barth, Aneta, and Aneta Barth. 2023. “The 3 Biggest Challenges in Analyzing ECG Signals – See If You Are Also Struggling with Them. (Part 2).” Cardiomatics. Krakow, Poland: Cardiomatics Sp. z o.o.

- Batra, A.; Jawa, V. 1975. “Classification of arrhythmia using conjunction of machine learning algorithms and ECG diagnostic criteria.” Train Journal 1: 1–7.
- Bordoloi, Monali, and Saroj Kumar Biswas. 2023. “Sentiment Analysis: A Survey on Design Framework, Applications and Future Scopes.” *Artificial Intelligence Review* 56 (11): 12505–60.
- Breiman, Leo. 2001. “Random Forests.” Machine Learning 45 (1): 5–32.
- Cambridge Heart Clinic. 2019. “Glossary of Terms.” Cambridge: Cambridge University Hospitals NHS Foundation Trust.
- Chen, Liang, Ze-Guang Han, Junhong Wang, und Chengjian Yang. 2022. “The Emerging Roles of Machine Learning in Cardiovascular Diseases: A Narrative Review.” Annals of Translational Medicine 10 (10): 611.
- Essa, Ehab, und Xianghua Xie. 2021. “Multi-Model Deep Learning Ensemble for ECG Heartbeat Arrhythmia Classification.” European Signal Processing Conference, January.
- Gao, J., Zhang, H., Lu, P., und Wang, Z. 2019. “An effective LSTM recurrent network to detect arrhythmia on imbalanced ECG dataset.” Journal of Healthcare Engineering 2019: 6320651.
- Goldberger, A., L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, und H. E. Stanley. 2000. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals.” Circulation 101 (23): e215–e220.

- Gupta, Aashuli, Arnob Banerjee, Disha Babaria, Kunal Lotlikar, und Hema D. Raut. 2021. “Prediction and Classification of Cardiac Arrhythmia.” In Advances in Intelligent Systems and Computing, 527–38. Stanford: University of Stanford.
- Hadaeghi, Fatemeh. 2019. “Reservoir Computing Models for Patient-Adaptable ECG Monitoring in Wearable Devices.” arXiv. Ithaca, NY: Cornell University.
- Hapke, Hannes, und Catherine Nelson. 2020. Building Machine Learning Pipelines. Sebastopol, CA: O’Reilly Media, Inc.
- Harris, C. R., K. Jarrod Millman, Stéfan J. Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. “Array Programming with NumPy.” Nature 585 (7825): 357–62.
- Hassan, Mehdi, Asmatullah Chaudhry, Asifullah Khan, und Muhammad Aksam Iftikhar. 2014. “Robust Information Gain Based Fuzzy C-Means Clustering and Classification of Carotid Artery Ultrasound Images.” Computer Methods and Programs in Biomedicine 113 (2): 593–609.
- Hunter, John D. 2007. “MatPlotLib: A 2D Graphics Environment.” Computing in Science and Engineering 9 (3): 90–95.
- Irfan, Saad, Nadeem Anjum, Turke Althobaiti, Abdullah Alhumaidi Alotaibi, Abdul Basit Siddiqui, und Naeem Ramzan. 2022. “Heartbeat Classification and Arrhythmia Detection Using a Multi-Model Deep-Learning Technique.” Sensors 22 (15): 5606.
- Javaid, Mohd, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, und Shanay Rab. 2022. “Significance of Machine Learning in Healthcare: Features, Pillars and Applications.” International Journal of Intelligent Networks 3 (January): 58–73. Beijing: KeAi Communications Co. Ltd.

- Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, und Tie-Yan Liu. 2017. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree.” Advances in Neural Information Processing Systems 30 (December): 3149–57. New York: Curran Associates.
- Keras-Team. 2015. “GitHub - Keras-Team/Keras: Deep Learning for Humans.” GitHub.
- Khanal, Raja Ram, et al. 2023. “Arrhythmias: Its Occurrence, Risk Factors, Therapy, and Prognosis in Acute Coronary Syndrome.” J Nepal Health Res Counc 21 (1): 8-14. Nepal: Nepal Health Research Council.
- Kıranyaz, Serkan, Türker İnce, and Moncef Gabbouj. 2016. “Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks.” IEEE Transactions on Biomedical Engineering 63 (3): 664–75.
- Lapid, Raz, und Moshe Sipper. 2022. “Evolution of Activation Functions for Deep Learning-Based Image Classification.” Proceedings of the Genetic and Evolutionary Computation Conference Companion, July.
- Luz, Eduardo, Thiago Nunes, Victor Hugo C. De Albuquerque, João Paulo Papa, und David Menotti. 2013. “ECG Arrhythmia Classification Based on Optimum-Path Forest.” Expert Systems With Applications 40 (9): 3561–73.
- Madani, Ali, Jia Rui Ong, Anshul Tibrewal, und Mohammad R. K. Mofrad. 2018. “Deep Echocardiography: Data-Efficient Supervised and Semi-Supervised Deep Learning towards Automated Diagnosis of Cardiac Disease.” Npj Digital Medicine 1 (1). London: Springer Nature Limited.

- Microsoft Research. 2023. “LightGBM.” January 24, 2023. Long Beach: Microsoft Redmond.
- Moody, G.B., und Roger G. Mark. 2001. “The Impact of the MIT-BIH Arrhythmia Database.” IEEE Engineering in Medicine and Biology Magazine 20 (3): 45–50.
- Moore, Alexander, und Max Bell. 2022. “XGBOOst, a Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study.” Clinical Medicine Insights: Cardiology 16 (January): 117954682211336. Auckland: Libertas Academica Ltd..
- Nasim, Amnah, und Yoon Sang Kim. 2022. “DE-PNN: Differential Evolution-Based Feature Optimization with Probabilistic Neural Network for Imbalanced Arrhythmia Classification.” Sensors 22 (12): 4450.
- Newman-Toker, David E., Zheyu Wang, Yuxin Zhu, Najlla Nassery, Ali S. Saber Tehrani, Adam C. Schaffer, C Winnie Yu-Moe, Gwendolyn Clemens, Mehdi Fanai, und Dana Siegal. 2020a. “Rate of Diagnostic Errors and Serious Misdiagnosis-Related Harms for Major Vascular Events, Infections, and Cancers: Toward a National Incidence Estimate Using the ‘Big Three.’” Diagnosis 8 (1): 67–84. Berlin: Walter de Gruyter.
- O'Riordan, Michael. 2022. “CVD Claimed 20 Million Lives in 2021, but Disease Burden Varies Globally.” tctmd.com, December 14, 2022. New York: TCTMD.
- Pandas development team. 2020. "pandas-dev/pandas: Pandas." Zenodo, February.
- Papa, João Paulo, Alexandre X. Falcão, Victor Hugo C. De Albuquerque, und João Manuel R. S. Tavares. 2012. “Efficient Supervised Optimum-Path Forest Classification for Large Datasets.” Pattern Recognition 45 (1): 512–20.

- Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." Advances in Neural Information Processing Systems 32: 8024–8035.

- Pedregosa et al. 2011. “Scikit-learn: Machine Learning in Python.” Journal of Machine Learning Research 12: 2825–2830. Brookline: Microtome Publishing.

- Rivo, Eduardo, Javier De La Fuente, Ángel Rivo, Eva García-Fontán, Miguel-Ángel Cañizares, und Pedro Gil. 2012. “Cross-Industry Standard Process for Data Mining Is Applicable to the Lung Cancer Surgery Domain, Improving Decision Making as Well as Knowledge and Quality Management.” Clinical & Translational Oncology 14 (1): 73–79. Springer Italia Srl.

- Saenz-Cogollo, Jose Francisco, und Maurizio Agelli. 2020. “Investigating Feature Selection and Random Forests for Inter-Patient Heartbeat Classification.” Algorithms 13 (4): 75.

- Sarfraz, Mohammad, Francis F. Li, und Asif Iqbal Khan. 2015. “Independent Component Analysis Methods to Improve Electrocardiogram Patterns Recognition in the Presence of Non-Trivial Artifacts.” Journal of Medical and Bioengineering 4 (3): 221–26. Lausanne: Springer International Publishing AG.

- Singh, N.; Singh, P. 2019. “Cardiac arrhythmia classification using machine learning techniques.” In Engineering Vibration, Communication and Information Processing. Singapore: Springer, 469–480.

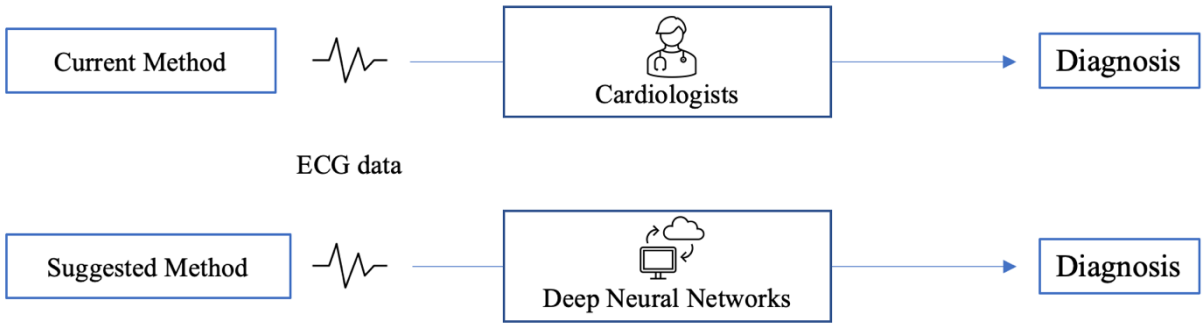
- Srinivasan, A. 2022. „Handbook of Research on Computer Vision and Image Processing in the Deep Learning Era.“ IGI Global. Hershey: IGI Global.

- Sturman, Oliver, et al. 2020. "Deep learning-based behavioral analysis reaches human accuracy and is capable of outperforming commercial solutions." Neuropsychopharmacology 45 (11). London: Nature Publishing Group.
- Taye, Mohammad Mustafa. 2023. “Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions.” Computers 12 (5): 91. Basel: MDPI.
- Waskom, Michael. 2021. “Seaborn: Statistical Data Visualization.” Journal of Open Source Software 6 (60): 3021.
- Wu, M.; Lu, Y.; Yang, W.; Wong, S.Y. 2021. “A study on arrhythmia via ECG signal classification using the convolutional neural network.” Frontiers in Computational Neuroscience 14: 564015. Lausanne: Frontiers Media S.A..
- Xie, Chen, McCullum, Lucas, Johnson, Alistair, Pollard, Tom, Gow, Brian, und Benjamin Moody. 2023. "Waveform Database Software Package (WFDB) for Python" (version 4.1.0) Massachusetts Institute of Technology: PhysioNet.

Appendix 1 – Previous / Proposed Model

The comparison of the current and the proposed model for the diagnosis of cardiac arrhythmias is illustrated below. The current method (shown above) is based on the analysis of ECG data by cardiologists, a process that is subject to subjective interpretation and can be time-consuming. In contrast, the proposed model (shown below) uses Deep Neural Networks (DNNs) to process the same ECG data, aiming to automate and standardize diagnosis. DNNs are able to recognize complex patterns in the data and could, therefore, increase accuracy and reduce diagnosis time. This model reflects the transition from a traditional, labor-intensive approach to a more efficient, technology-driven process and offers the potential to revolutionize the treatment of cardiac arrhythmias.

FIGURE 3: Overview of Proposed Process



Appendix 2 – Existing Literature

A compilation of selected relevant results from previous research is presented below. This overview is tabulated and highlights the findings of other researchers using the MIT-BIH dataset. The goal of this presentation is to provide a comprehensive view of the current state of research and a benchmarking reference to the present research results. It should be noted that the results are presented only in excerpts, each to consider similar methodological approaches to allow for a high degree of comparability.

TABLE 6 – Literature Review Overview

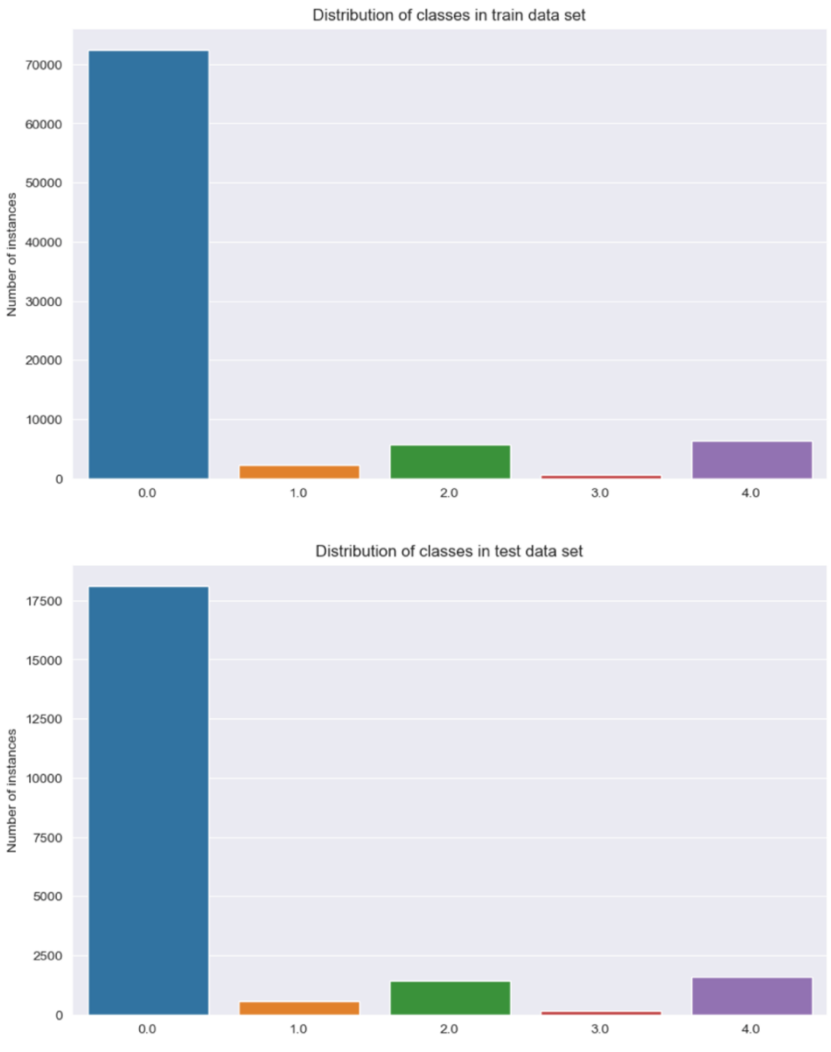
Publisher	Algorithm Technique	Accuracy
Hassan et al. 2014 (*)	FCM PNN	97.54 %
Nasim and Kim 2022 (*)	DE-PNN	99.33 %
Irfan et al. 2022 (*)	Multi-Model DL	98.41 %
Gupta et al. 2021 (*)	Ensemble (NB, RF, SVM)	77.40 %
Saenz-Cogollo and Agelli 2020 (*)	RF (Feature Selection)	96.14 %
Luz et al. 2013 (*)	OPF	90.10 %
Batra and Jawa 1975	GB-SVM	84.82 %
Singh and Singh 2019	TFFS (RF+BFS)	85.58 %
Alfaras, Soriano, and Ortín 2019	ESN	98.60 %
Gao et al. 2019	LSTM	95.80 %
Wu et al. 2021	DCNN (TFCV)	97.40 %
Hadaeghia, F. 2019	RNN + ESN	99.11 %
Proposed Work	**CNN**	**98.53 %**

* Entries marked with an asterisk represent sources discussed in the Literature Review of this paper.

Appendix 3 – Variable Distribution

The analysis of the data distribution according to AAAMI/ANSI standards for ECG classifications shows a clear imbalance: normal heartbeats (class 0) dominate with 90,631 cases, while supraventricular ectopic beats (class 1) and fusion beats (class 3) are strongly underrepresented with only 2,781 and 803 cases respectively. This could be due to a lower probability of occurrence or underreporting in the data sets. Ventricular ectopic beats (class 2) and beats of undetermined type (class 4) have a medium frequency with 7,239 and 8,043 cases respectively. This distribution underlines the necessity of data balancing/ weighting for the training of classification algorithms, which is explained in more detail in Chapter 3.1.

FIGURE 4: Variable Distribution



Appendix 4 – Results of LSTM

In addition to the methods discussed in Chapter 3.4, further methods were tested in order to develop a comprehensive model. The aim was to evaluate a wide range of methods to enable a sound final assessment that takes into account common practices. The following results reflect tests of these methods. The same model configuration was used as in that chapter, but with a modification in the last layer: instead of a Conv1D, an LSTM was used. This adaptation was based on findings from the literature review, which revealed the increasing use of this component in current models. The model developed aims to reflect the state of the art and therefore integrates these analyses. The corresponding results are presented below.

TABLE 7 - Results of LSTM

Classes	Accuracy	Precision	Recall	F1	MCC
N	0.9881	0.9933	0.9923	0.9928	0.9585
S	0.9920	0.8348	0.8543	0.8444	0.8404
V	0.9953	0.9746	0.9544	0.9644	0.9620
F	0.9961	0.6954	0.8457	0.7632	0.7650
Q	0.9990	0.9932	0.9932	0.9932	0.9926

The results show that the model, in complementary addition to the self-developed model, is characterized by above-average precision values in the classification. Nevertheless, it was observed that better average performance metrics were achieved without the use of Long Short-Term Memory. Specifically, the average F1-score and Matthews Correlation Coefficient (MCC) of the LSTM model were 91.16% and 90.37%, respectively, while the model without LSTM achieved values of 92.30% and 91.49%, respectively. As a result, due to the lack of performance improvement, it was decided not to use an LSTM in the optimization phase.

Appendix 5 – Performance Metrics

In order to quantify the performance of both the individual prototypes and the further developed convolutional neural networks (CNN), the following metrics were chosen: Accuracy, Precision, Recall, F1 and Matthews Correlation Coefficient.

A corresponding overview of the calculation of these can be found below:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{\text{Correct predictions}}{\text{Total data points}}$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{\text{Correctly predicted positive}}{\text{All predicted positive}}$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{\text{Correctly predicted positive}}{\text{All real positives}}$$

$$\text{F1-Score} = 2 \text{ x } \frac{\text{precision} \cdot \text{recall}}{\text{precision}+\text{recall}}$$

$$\text{Matthews Correlation Coefficient} = \frac{TN \cdot TP - FN \cdot FP}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Appendix 6 – Confusion Matrixes Prototype Models

The following appendix contains additional information that primarily serves to deepen and broaden understanding. The content presented here is discussed in more detail in the *Analysis and Discussion* section. It should be noted that the data and results presented are merely representative examples that should be interpreted in the context of the entire thesis. Furthermore, it should be mentioned that slight variability in the results may occur during the modeling and training of the models, depending on specific runs. Despite these potential fluctuations, this thesis is based on the results and findings presented here, which form the basis for the conclusions and recommendations developed

FIGURE 5: Confusion Matrix Stacking Model (Prototype)

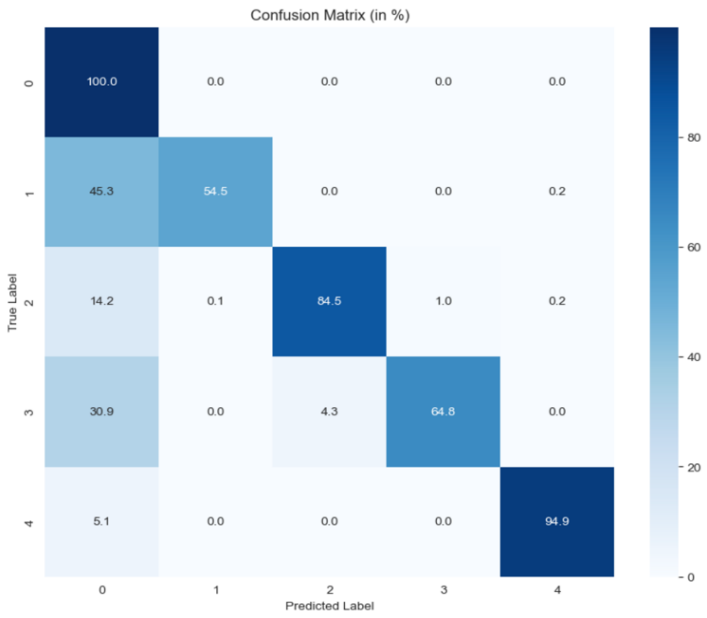


FIGURE 6: Confusion Matrix LightGBM Model (Prototype)

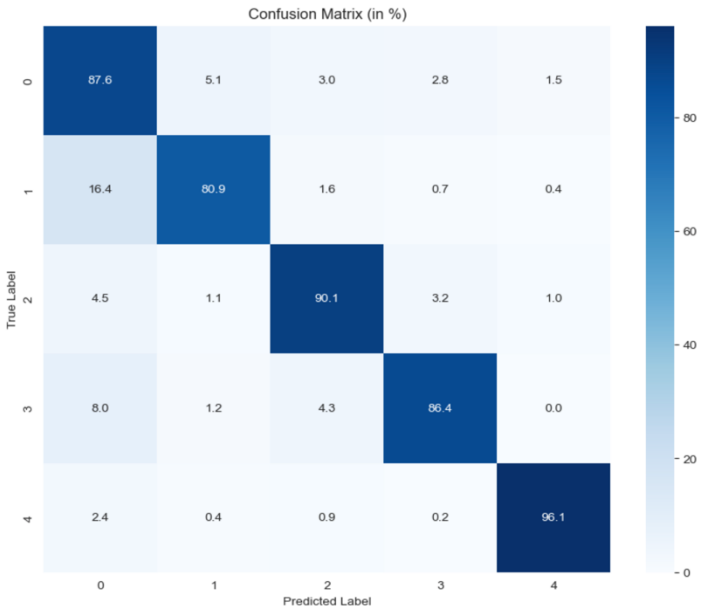
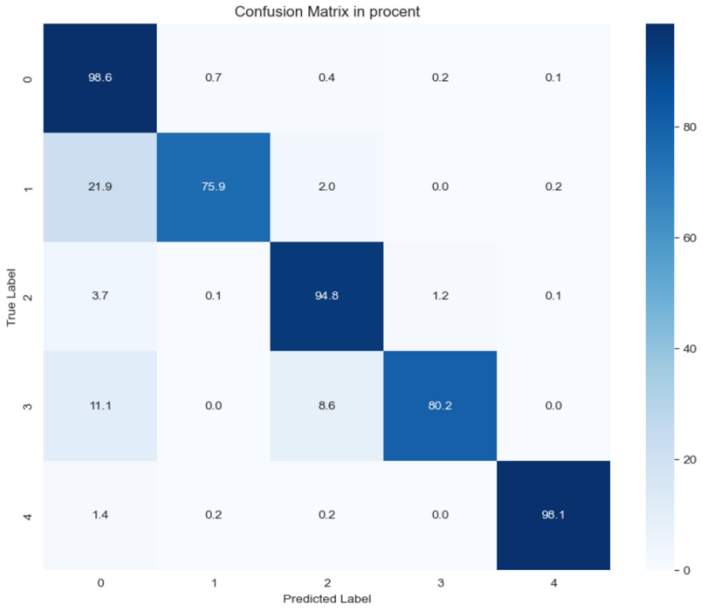


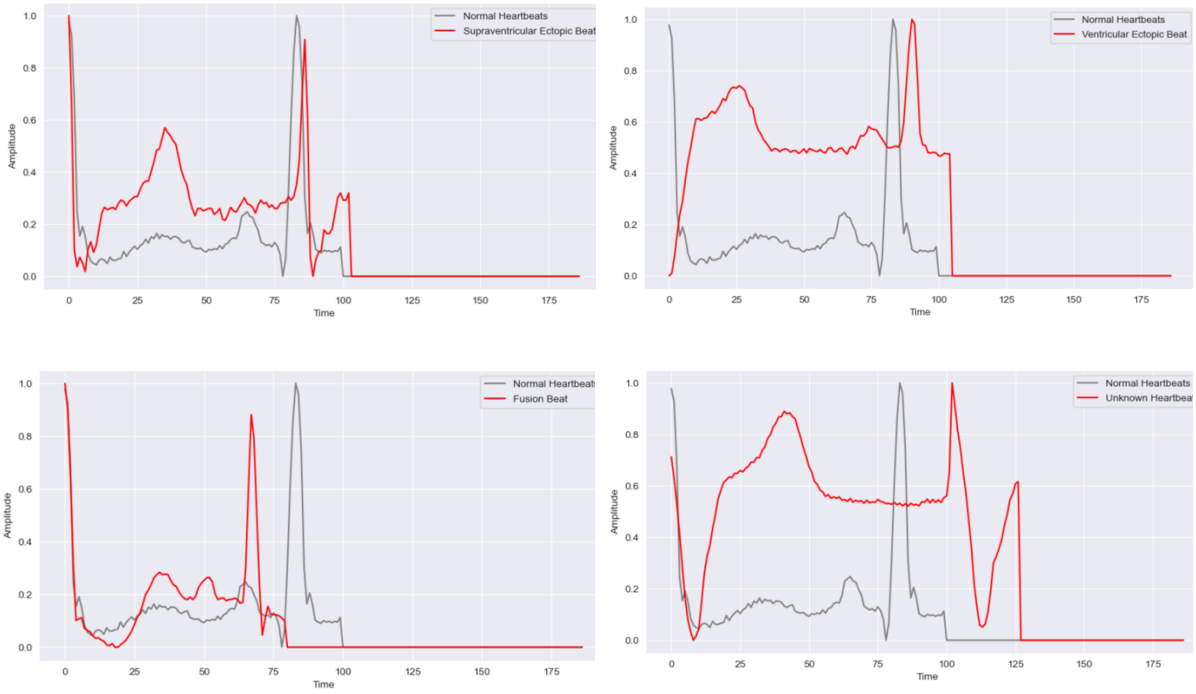
FIGURE 7: Confusion Matrix Convolutional Neural Network (Prototype)



Appendix 7 – Arrhythmia cases

The visual representation of ECG signals in the figure below illustrates the characteristic differences and similarities between normal and abnormal heartbeats. Normal heartbeats show a consistent, recognizable pattern characterized by a QRS complex and T wave. Supraventricular and ventricular ectopic beats show marked variations in the morphology of the QRS complex, indicating their aberrant site of origin in the heart. Fusion beats, which result from the combination of a normal and an ectopic beat, show a mixed morphology, which is reflected in a partial overlapping of the signal characteristics. Unknown heart beats, on the other hand, show a strongly varying signal shape that deviates from the norm and does not allow a clear assignment to the typical classes. This visualization underlines the importance of careful analysis of ECG signal morphology for accurate classification and the need for advanced algorithms to distinguish complex patterns in medical diagnostics.

FIGURE 6: ECG Arrhythmia cases



Appendix 8 – Results of Dropout Rates Comparison (in %)

Based on the findings from Chapter 4.3, the hypothesis was formulated that the model could be prone to overfitting. As a countermeasure, an increase in the dropout rate was proposed in order to improve the generalization capability. Experiments with variable dropout rates confirmed the effectiveness of this method to increase the robustness of the model. As shown in Table 8, a 30% dropout rate achieved an average accuracy of 99.15%, a precision of 84.72% and an F1 score of 87.55%. Reducing the dropout rate to 20% resulted in a slight improvement in accuracy to 99.34%, precision to 90.06% and F1 score to 91.43%.

TABLE 8 - Results of 30% Dropout / 20% Dropout rate comparison (in %)

Using 30% dropout rate						Using 20% dropout rate				
Classes	Acc.	Prec.	Rec.	F1	MCC	Acc.	Prec.	Rec.	F1	MCC
N	98.16	99.34	98.43	98.88	93.69	98.64	99.33	99.02	99.18	95.26
S	99.26	85.79	84.71	85.25	84.87	99.19	84.14	83.99	84.07	83.65
V	99.32	94.28	95.58	94.92	94.56	99.36	94.07	96.34	95.19	94.85
F	99.12	45.08	87.65	59.54	62.51	99.67	73.44	87.04	79.66	79.79
Q	99.88	99.13	99.19	99.16	99.09	99.86	99.31	98.82	99.06	98.99
Average	99.15	84.72	93.11	87.55	86.94	99.34	90.06	93.04	91.43	90.51

The optimized model with a 10% dropout rate outperformed these values by achieving an average accuracy of 99.45%, a precision of 92.28%, a recall rate of 92.34%, an F1 score of 92.30% and a Matthews Correlation Coefficient (MCC) of 91.49%. These results underline the superiority of our approach and confirm the effectiveness of an adjusted dropout rate to avoid overfitting without compromising the pattern recognition ability of the model on new data.