2025 IMEKO Joint Conference
TC8 - TC11 - TC25
Torino, Italy, September 14-17, 2025

The Impact of Data Quality on ML Diagnostic Models: Ensuring Reliability in Medical AI

Bartlomiej Cieszynski[1] and Joao Gregorio[1]

[1]*Informatics, Data Science Department, National Physical Laboratory, Teddington, TW11 0LW, London, United Kingdom, bartlomiej.cieszynski@npl.co.uk*
[2]*Informatics, Data Science Department, National Physical Laboratory, Glasgow G1 1RD, United Kingdom, joao.gregorio@npl.co.uk*

***Abstract* – The application of artificial intelligence and machine learning algorithms in healthcare has grown exponentially in recent years, offering benefits such as improving diagnostic accuracy, efficiency, and addressing challenges such as interpretative bias or the increasing volume of patient data. However, ensuring the reliability of such models necessitates robust evaluations of their sensitivity to data quality – the extent to which data meets required standards. This study explores how variations in data accuracy, precision, and completeness affect an algorithms' ability to correctly classify electrocardiogram results. The models analysed, namely K-Nearest Neighbours, Random Forest, Artificial Neural Networks, and Convolutional Neural networks were selected due to their prevalence in electrocardiogram classification tasks. Using the PhysioNet's MIT-BIH Arrhythmia Dataset, the study classifies five types of beats defined by the AAMI EC57. To simulate varying data quality, the dataset has undergone systematic degradation through the addition of noise, rounding of numerical, and removing data features. The re-evaluation of model performance, quantified by model accuracy, precision, and recall has highlighted the effects of data quality on diagnostic outcomes, providing insights into the robustness of models under suboptimal conditions. By examining these critical factors, the study aims to inform the development of more reliable machine learning diagnostic systems, raising awareness of the importance of data integrity in medical applications. The study has shown that model performance is most sensitive to accuracy, completeness and precision in descending order; with accuracy showing greatest reduction in model predictions. It has also displayed that model sensitivity is correlated with class population, as low represented classes have yielded greater deviation under the application of data degradation.**

I. INTRODUCTION

The integration of Machine Learning (ML) and Artificial Intelligence (AI) across diverse industries has grown significantly in recent years, revolutionising fields such as healthcare, environmental science, and pharmaceuticals, as well as transforming operations in finance, retail, and manufacturing [2, 21, 20, 22]. However, the adoption of AI in certain domains, such as healthcare, where decisions made on erroneous AI predictions could result in significant financial losses or safety risks, requires a rigorous procedure to analyse the robustness and trustworthiness of performance and outputs from AI models.

With current challenges in the national healthcare system, AI offers a promising solution to improving its services. By enabling faster and more accurate diagnostics, AI models can alleviate the burden on healthcare systems, improve decision-making, and reduce waiting times by helping clinicians manage the rising volume and complexity of patient information. The ability of AI models to learn from vast datasets allows it to identify patterns and make predictions that may not be immediately apparent to human practitioners. For AI models to be reliable and resilient in healthcare applications, validation, verification and uncertainty quantification are necessary to ensure its robustness, effectiveness, and safety under diverse clinical conditions.

There is a direct correlation between the quality of the input and output of the ML model [5]. This study examines how variations in Data Quality (DQ) - the extent to which data meet the specifications established by an organisation responsible for developing a product [13, 9] - influence the performance of ML algorithms in accurately diagnosing medical conditions based on Electrocardiogram (ECG) data. The models evaluated in this work include K-Nearest Neighbour (KNN), Random Forest (RF), Artificial Neural Network (ANN), Convolutional Neural Network (CNN), One-vs-Rest Logistic Regression (OVR), and Decision Tree (DS) algorithms. Examining how DQ dimensions impact ML model performance enhances our understanding of these models, builds trust, and provides insights into their application under varying conditions.

The exploration of ML methods using ECG data in diagnostics has been an active field of research in recent years, with many studies evaluating the prediction accuracy of various models [14, 8, 3]. These studies focus on model optimisation, feature extraction, and data handling. Analysis of the effects of DQ and the reliability of these

models garnered less attention. This work aims to support the development of reliable, robust, and resilient AI and ML in medical diagnostics, emphasising the importance of DQ in the training process. To achieve this, the available ECG dataset has undergone data degradation, specifically, a reduction in data accuracy, precision, and completeness. This has been done through application of noise, rounding of values, and removal of features (dataset columns).

II. MATERIALS AND METHODS

A. Dataset

This study uses the the MIT-BIH Arrhythmia ECG Heartbeat Categorisation Dataset, available on Kaggle [4] to develop a case study investigating the relationship between model performance and data quality. The dataset is derived from PhysioNet's MIT-BIH Arrhythmia Dataset [15] and comprises 109,446 heartbeat samples, sampled at 125 Hz. It is defined in accordance with the AAMI EC57 standard [1, 10], and is partitioned into five beat classes: class N (0), 72 471 samples, includes normal sinus rhythm, left and right bundle branch block (LBBB, RBBB), atrial escape and nodal escape; class S (1), 2 223 samples, includes supraventricular premature, aberrated atrial premature, nodal (junctional) premature, atrial premature and ectopic rhythms; class V (2), 5 788 samples, includes premature ventricular contractions (PVC) and ventricular escape beats; class F (3), 641 samples, consists of fusions of ventricular and normal beats; and lastly class Q (4), 6 431 samples, covers unclassified beats and noise.

The data has been imported in the form of a Comma-Separated-Values (CSV) file, with the last column of the data denoting the class of the sample. The data additionally provides both training and testing samples for ML purposes, and has been standardised such that all values are between 0 and 1. However, a limitation of our methodology is that it does not address the class imbalance present in the original dataset. Maintaining class imbalance, although may result in lowered performance of model for less represented classes, will allow to establish any relationship between class representation and DQ.

B. ML Models

The models have been chosen based on their prevalence in ECG classification studies [10, 2, 14, 21, 3], as well as their ease of implementation. All of the models, apart from the CNN, have been implemented through the `scikit-learn` Python library, while the CNN has been developed with the use of `Pytorch`.

B..1 One-Vs-Rest Logistic Regression

The OVR algorithm is a multi-class adaptation of logistic regression [17, 7], designed to estimate the probability of an outcome being true given a predictor. While standard logistic regression is suited for binary classification, OVR extends this approach by training multiple classifiers simultaneously. Each classifier calculates the probability of a sample belonging to one specific class versus all other classes by maximising a likelihood function. The closer the output of the function is to 1, the greater the probability that the sample belongs to the target class over the alternatives. This can be interpreted as multiple binary classification predictions, where each individual class is compared with the rest as collective, this is further reflected in the name of the algorithm "One-Vs-Rest',.

For this study, the OVR model was implemented using the `OneVsRestClassifier` function from the `scikit-learn` library. The underlying classifier was `LogisticRegression`, configured with a maximum of 1,000 iterations, the limited memory Broyden-Fletcher-Goldfarb-Shanno solver, and a regularisation strength of 10.

B..2 Decision Tree

DT models operate by recursively splitting the dataset into subsets based on features or attributes, forming a tree-like structure of decision nodes and leaves. Each decision node represents a condition on a feature, while each leaf corresponds to a predicted class or value. The splits are determined by maximising a criterion such as the Gini impurity or information gain, which measures the quality of the split. For the purpose of this study, the model has been initialised using default parameters set by the `scikit-learn DecisionTreeClassifier` function with a random state of 42.

B..3 Random Forest

The RF model is an ensemble learning method that builds decision trees during training and combines their results to improve the accuracy and robustness of predictions. It follows a process called bootstrap aggregating, which randomly selects data samples with replacement; a process which allows for re-selection of points with each iteration [23].

For this study, the RF model was implemented using the `RandomForestClassifier` function from the `scikit-learn` library. The model was optimised with the parameters: `max_depth`: None; `min_samples_leaf`: 1; `min_samples_split`: 2; `n_estimators`: 300; `random_state`: 42.

B..4 K-Nearest Neighbours

KNN is an instance-based algorithm that makes predictions by measuring the similarity between data points [11]. It does not involve an explicit training phase; instead,

it memorises the entire dataset and uses it during inference. Predictions are made by comparing a new data point with stored examples through distance metrics such as the Euclidean, Manhattan, Minkowski, or Hamming distance [16]. The algorithm assigns the new data point to the class most common among its nearest neighbours, determined by the chosen distance metric.

The KNN algorithm was implemented in Python using the `KNeighborsClassifier` function from the `scikit-learn` library. The model was configured with three nearest neighbours, compared using the Manhattan distance metric and weighted by distance, before being applied to the ECG data.

B..5 Artificial Neural Network

The ANN in this study has been implemented in the form of a Multi-Layer Perceptron (MLP) model provided by the `scikit-learn` library. The MLP is a supervised algorithm which uses feature based neurons, where each neuron transforms its input from the previous layer through a weighted and biased linear summation followed by an activation function. The model is trained using iterative backpropagation algorithm, which involves evaluating the error through a loss function, in this case the Cross-Entropy loss - a negative log-likelihood quantifying the deviation in a prediction probability distribution and the true distribution [6, 12], and adjusting the weights to minimise this error. For multi-class classification tasks, a softmax function is applied to the output layer of the ANN to produce probability distributions over the classes [19].

The MLP has been applied with a network structure of 128, 64, 32 neurons, the ReLU activation function, 300 maximum iterations, and a random state of 42. All other parameters have been set to default settings of the `scikit-learn` `MLPClassifier` function.

B..6 Convolutional Neural Network

A CNN differs from a traditional ANN primarily in its architecture and how it processes data. While ANNs employ fully connected layers where each neuron is connected to every neuron in the subsequent layer, CNNs use a hierarchical structure consisting of convolutional layers, pooling layers - which generally reduce the dimensionality of data - and fully connected layers. Convolutional layers apply kernels (filters) that slide across the input data, performing element-wise multiplications to extract spatial features such as patterns. These kernels enable CNNs to learn local and spatial hierarchies of features. Pooling layers, often using operations like max-pooling, reduce the spatial dimensions of the feature maps, thereby lowering computational complexity and mitigating the risk of overfitting by summarising feature information. Dropout layers are used as regularizers to reduce overfitting by randomly disabling a fraction of neurons during training. Finally, the extracted features are passed through fully connected layers, and the network's output is computed using an activation function, such as softmax, which transforms the results into probabilities corresponding to the input's classification into predefined categories.

For the purpose of this study, a CNN with a random seed of 30, two convolutional layers (32 and 64 filters of size 5 with stride 1 and padding 2), two pooling layers (kernel of size 2 and stride 2), a ReLu activation function, a 50% dropout layer, and two fully connected layers have been used. The network was created in python through the use of the `Pytorch`.

C. Data Quality Dimensions

Data accuracy, precision, and completeness have been identified as key DQ dimensions, where, accuracy is "the degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use", precision is "the degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use", and completeness is "the degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use" [13].

To determine the impact of these dimensions on the original MIT-BIH dataset, data degradations scripts have been developed to investigate model performance. The original dataset was used to establish a baseline for assessing data quality. Hence, to simulate the effect of the above dimensions the data has undergone the following.

To reduce data accuracy, varying intensities of Gaussian noise have been applied. A noise array, controlled by the magnitude of its standard deviation, has been generated in the form of a normal distribution. This array has then been added onto the original training dataset to mimic the effects of lowered accuracy.

To reduce data precision, the original dataset has been recreated with rounding to a varying number of decimal places. Specifically, the data has been rounded from 16 to 13, 10, 7, and 4 decimal places.

To reduce data completeness, the original data has been downsampled by removal of ECG features. To achieve this, data features were systematically removed by retaining only every second, third, or fifth column in each dataset facilitating a controlled reduction in features.

D. Model Evaluation Metrics

To assess the performance of the models, the results have been summarised by the `scikit-learn` `classification_report` function which outputs a

score of model precision, recall, f1-score, support, and overall accuracies. In this study, three key metrics were selected for model evaluation: precision, recall, and overall accuracy. Precision measures the proportion of true positives among all positive predictions, indicating how often the model correctly identifies each class when making a positive prediction. Recall represents the proportion of actual positive instances correctly identified by the model. Overall accuracy quantifies the proportion of correct predictions relative to the total number of predictions made [18].

III. RESULTS

The application of methods discussed in Section ii. has resulted in multiple classification reports which allowed to quantify the model performance. The OVR and DT algorithms exhibited poor classification performance on the original dataset, with precision and recall values ranging from 10% to 60%. Due to their limited effectiveness in multinomial classification, these models were excluded from further analysis.

Figure 1 presents the overall accuracy of each remaining model across different datasets. The x-axis represents the evaluated models, while the y-axis denotes their accuracy. The results indicate consistently high classification performance, with accuracy ranging from 95% to 98.5%, regardless of input data quality. Conversely, the results indicate that the model performance was most sensitive to data accuracy, as the introduction of noise to data led to the most significant decline in accuracy.

Figure 2 presents subplots for each model, illustrating the variation in precision and recall across degraded datasets for each class. These results show something that overall accuracy did not account for: that prediction performance for Classes 1 and 3 is substantially lower than for other classes, which are the least represented classes by volume in the dataset. This result highlights the sensitivity of the ML models to class representation.

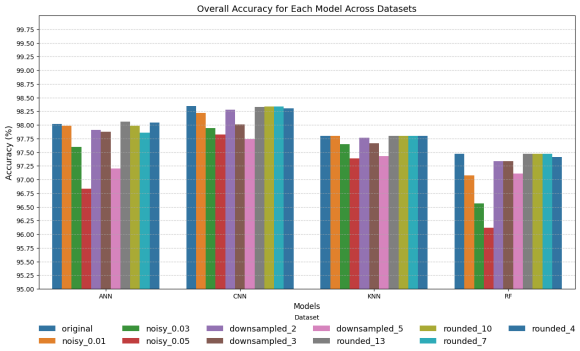

Fig. 1. Bar chart displaying model performance for each dataset based on overall accuracy of predictions.

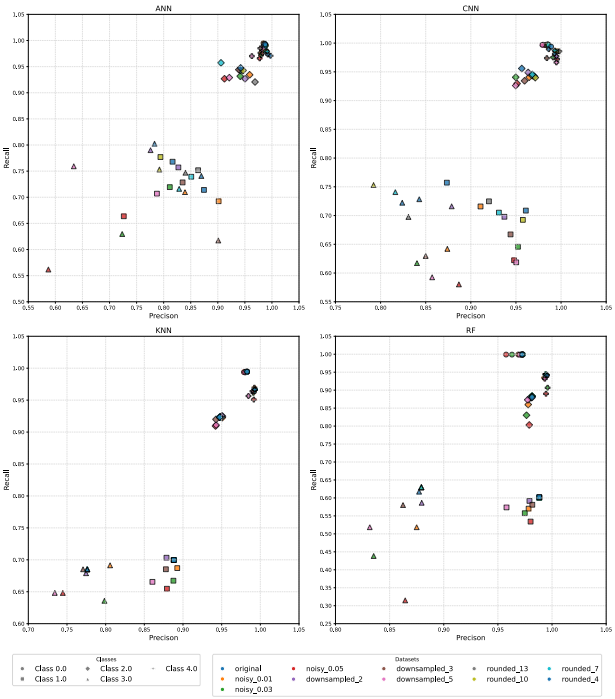

Fig. 2. Full model comparison based on precision versus recall. Marker shapes indicate different classes, colors represent datasets.

IV. DISCUSSION

The OVR and DT models were excluded from further analysis due to their poor initial performance. While their overall accuracy was acceptable, at 90.7% and 95.2%, respectively, both models, particularly OVR, exhibited significant sensitivity to data volume and class representation. The OVR model yielded recall values of 15% and 21% for classes 3 and 1, respectively, and precision values of 65% and 77% for classes 2 and 1. Although DT outperformed OVR, its recall and precision values for classes 3 and 1 remained suboptimal at 57%, 63%, 56%, and 62%, respectively. Given their comparatively poor performance on key metrics without further optimisation, these models have been excluded from further analysis. This decision is based on the principle that extensive optimisation could obscure the effects of data quality degradation, as a well-tuned model may compensate for lower-quality inputs, making it more difficult to assess the true sensitivity of model performance to variations in data quality.

A. *Model Accuracy*

The results of this dataset indicate that ML models are most sensitive to degradation in regards to data accuracy as each model has experienced reduction in performance by using noisy datasets for training, in contrast to the use of the original dataset. The ANN and RF models in particular

have shown biggest sensitivity with an approximate 1% reduction. Although CNN and KNN have experienced lower reduction in model accuracy they still show a steady decrease in performance from the introduction of noise to the training data. However, despite the effects of noise yielding reduction in model accuracy in the order of a percent, this does not truly represent the effect of data degradation on model predictions. Instead, this is a results of bias introduced by a combined accuracy, which negates the impact of class population and imbalance. Class 0 with approximately 70,000 samples dominates the dataset and yields correct predictions between 95% and 99% of the time, which significantly overshadows the reduction in prediction accuracy for less represented classes.

Although the effects of degrading data completeness have not matched or exceeded the effects resulting from the degradation of accuracy, the results indicate that completeness also affects model performance to a noticeable degree. Each model experienced a reduction in accuracy caused by the reduction of features, with more significant losses for the ANN and CNN models in comparison to the KNN and RF algorithms. Additionally, RF showed no change between the downsampled_2 and downsampled_3 datasets which indicates either low sensitivity to data completeness or improper model optimisation.

Figure 1 shows the lack of model sensitivity to degradation of data precision. Each model displayed little to no reduction in classification accuracy whereas, the ANN model displayed unexpected behaviour. The initial reduction of precision from 16 to 13 decimal places has increased the model performance, while the reduction from 16 to 10 has yielded approximately the same result. Rounding of values further to four decimal places has resulted in an increase of model accuracy when compared to the same model training with the original data. A possible explanation for this phenomena can be found in the use of data smoothing - reduction of "sharp" data characteristics - as a common preprocessing technique to refine ML training to avoid overfitting.

The observations discussed above imply that the models evaluated are most sensitive to degradation of accuracy, completeness, and precision in descending order. This may be explained by considering how each data quality metric actually affects the data. Addition of noise, affecting dataset accuracy, resulted in the highest reduction in model performance, caused by the introduction of random variation to dataset values. Considering the models use the training data to observe patterns and trends in features, altering these patterns can therefore reduce the models capability. Both completeness and precision may not explicitly affect the data trends due to the systematic application of data degradation. Rounding of all values and removal of the same features in variable amounts may not explicitly affect data trends in comparison to the the addition of random noise to training data. However, completeness introduces another difficulty, although it may not directly affect data patterns numerically, removal of features may result in a similar outcome. This can therefore lead to the observations made in Figure 1. Depending on the dataset resolution, systematic removal of features may not result in loss of information due to smoothing effects.

B. Model Precision & Recall

Figure 2 allows for further evaluation of models on the scale of individual classes and datasets; allowing for a more robust review of model performance. It can be observed that classes 0, 2, and 4 perform substantially better than classes 1 and 3 on the basis of model precision and recall, implying the sensitivity of the model to sample size and class representation.

The MIT-BIH dataset is not balanced as the number of available data samples for each class varies significantly. Class 1 which is most represented in the data contains approximately 70,000 data points. The predictions for such a well-populated class have been successful as all models yield results above 97% for precision and recall. As the class representation decreases, so does the model predictive capabilities. This can be seen from interpreting the performance of class 3 predictions, which is the least represented class with only 641 samples, as the models have obtained precision and recall scores in the range between 30% to 80%.

The results indicate that the evaluated ML models, for classification tasks, are more impacted by class representation and data volume, rather than the quality of the training dataset. Degradations in data quality resulted in loss of performance on the scale of a percent, whereas, the sample size can result in reduction of almost 70% in the case of RF predictions. The contrary however, can be observed from the performance of the RF model, where class 2 predictions display a drop in recall by approximately 7% and class 1 precision reduced by approximately two percent; indicating a higher sensitivity of the model to quality of training data even on well-populated classes. Despite the clear effect of noise on the model, other models do not display the same effects.

Although the effect of sample size are very clear and apparent, conclusions on the effects of data degradation combined with low populated classes are ambiguous. This is due to the vast spread of results for the less represented classes in all models. Classes 0, 2, and 4 present to have little deviation between the datasets as the points are all within a small range. The range of results for completeness and precision may be further validated due to the effects mimicking data preprocessing, such as smoothing, where rounding or reduction of features may reduce overfitting and hence increase model performance. Alternatively, class 1 and 3 show a wide spread of results for both

precision and recall in all models. This may suggest that models are more sensitive to data quality metrics in the absence of a well-populated sample. This once again appears intuitive due to the nature of the algorithms. Observation of patterns in the data which allow for the classification may be limited in less populated samples, this may be further impacted by noise, missing features or rounding of values. This is further supported by the lower deviation of results for the KNN model, where predictions are made by direct calculation rather than indirect learning. Hence, the currently displayed results do not allow for conclusive comparisons of the effects of data degradation combined with class representation.

C. Model Comparison

Figure 2 suggests that the CNN and KNN models outperform the ANN and RF algorithms based on the distribution of precision and recall values. However, the nature of this research has not allowed for an explicit 'better model' comparison. The aim was not optimising models to obtain best results, but analysing the effects of training data on the performance of each model.

Alternatively, the sensitivity of each model to data quality can be discussed. The ANN shows overall high performance, with a loss of prediction accuracy for less represented classes. Figure 2 also shows the sensitivity of the model to data quality. For classes 1, 2, and 3, an almost linear relationship can be observed from the performance of noisy data, with the noisy_0.05 dataset resulting in the lowest precision and recall values. Data accuracy however, is the only metric with a visible correlation. All other datasets display a random nature affecting performance.

A similar conclusion can be made on the CNN model, where results for class 1 and 3 show a clear reduction in model precision under application of noise and reduction of features. However, the same cannot be said for model recall as the noisy_0.05 dataset scored higher than the noisy_0.01. This negates any conclusions on the effect of data accuracy and completeness on model recall. Overall, the model reinforces the sensitivity of class sample on predictions as the higher populated classes show the lowest result deviation out of all the models.

Figure 1 also suggests that the KNN model is the least affected by data quality, as evidenced by its consistent performance despite variations in precision, accuracy, and completeness. This may be a result of the nature of the algorithm, as the KNN model does not undergo a direct learning phase. Instead, it "memorises" the dataset and evaluates the Manhattan metric for each sample. Such a direct comparison alters the outcome from a standard prediction to a more robust calculation which can therefore reduce sensitivity to variation in data patterns. However, despite the robust method of this model, it may not account for key patterns between data features which help categorise samples, as seen from a higher overall performance of the CNN model.

Based on observations from Figure 2 the model shows a very clear correlation between data quality and model recall. Classes 1 and 3 show little variation in precision results but a clear reduction of recall under the degradation of accuracy and completeness. Additionally, the RF algorithm is the only model which shows variation in class 0 results, with the noisy_0.03 and noisy_0.05 datasets reducing model precision. This indicates the RF model is most sensitive to degradation of data quality as even such a well represented class suffers from the use of lower quality training data as input. This could be a result of the nature of the algorithm as, unlike neural networks, the RF evaluates a likelihood mathematically. This suggests that deviation in data values could lead to more explicit variations in the Gini impurity resulting in greater sensitivity to training dataset quality.

V. CONCLUSIONS

This study investigated a range of ML models for supporting medical diagnosis by using ECG data for classification tasks. In particular, the effects of data accuracy, precision and completeness have been considered as key data quality metrics affecting the performance of models such as the MLP ANN, CNN, KNN, and RF. The models have shown different levels of sensitivity to data quality, and the effects seem to directly correlate with the sample size of classes the models try to classify. Lower represented classes have shown a clear dependence on data quality, whereas the well-populated classes showed lower dependence on data quality when compared with their sparsely populated counterparts. This has been reflected in the results, where models such as the ANN and CNN yielded a substantial range in prediction precision and recall, approximately 55% to 90% and 55% to 80%, respectively, for the least represented class 3. The ANN model also displays an almost linear correlation between dataset accuracy and predictive performance for class 3. Additional clear correlations between dataset quality and predictive performance may be observed for lower represented classes in the CNN, KNN, and RF models. These correlations however, apply to either precision or recall individually, with less visual correlation on combined performance. Meanwhile, well-populated classes, such as class 0 and 4, displayed sub 5% ranges in both precision and recall for all models, suggesting class representation as the dominating factor of model predictive capabilities. Despite the comparison of model performance, this study was not tailored for direct determination of a 'better' model due to the lack of emphasis on model optimisation, and the focus on training dataset quality.

VI. CITATIONS AND REFERENCES

[1] AAMI. *Testing and Reporting Performance Results of Cardiac Rhythm and ST Segment Measurement Algorithms (EC57)*. Standard. Available at: https://www.aami.org/. Arlington, VA, USA: Association for the Advancement of Medical Instrumentation (AAMI), 2012.

[2] Ahmed Al Kuwaiti et al. "A review of the role of artificial intelligence in healthcare". In: *Journal of personalized medicine* 13.6 (2023), p. 951.

[3] Yehualashet Megersa Ayano et al. "Interpretable machine learning techniques in ECG-based heart disease classification: a systematic review". In: *Diagnostics* 13.1 (2022), p. 111.

[4] Shayan Fazeli. *ECG Heartbeat Categorization Dataset*. `https://www.kaggle.com/datasets/shayanfazeli/heartbeat`. Accessed: 2024-05-13. 2019.

[5] R Stuart Geiger et al. ""Garbage In, Garbage Out" Revisited: What Do Machine Learning Application Papers Report About Human-Labeled Training Data?" In: *arXiv preprint arXiv:2107.02278* (2021).

[6] Ian Goodfellow. *Deep learning*. Vol. 196. MIT press, 2016.

[7] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[8] Essam H Houssein, Moataz Kilany, and Aboul Ella Hassanien. "ECG signals classification: a review". In: *International Journal of Intelligent Engineering Informatics* 5.4 (2017), pp. 376–396.

[9] ISO. *ISO/IEC 25012:2008; Software Engineering—Software Product Quality Requirements and Evaluation (SQuaRE)—Data Quality Model*. Technical Report. Geneva, Switzerland: International Organization for Standardization, 2008.

[10] Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. "Ecg heartbeat classification: A deep transferable representation". In: *2018 IEEE international conference on healthcare informatics (ICHI)*. IEEE. 2018, pp. 443–444.

[11] Oliver Kramer and Oliver Kramer. "K-nearest neighbors". In: *Dimensionality reduction with unsupervised nearest neighbors* (2013), pp. 13–23.

[12] Anqi Mao, Mehryar Mohri, and Yutao Zhong. "Cross-entropy loss functions: Theoretical analysis and applications". In: *International conference on Machine learning*. PMLR. 2023, pp. 23803–23828.

[13] Russell Miller et al. "A Framework for Current and New Data Quality Dimensions: An Overview". In: *Data* 9.12 (2024), p. 151.

[14] Ana Mincholé et al. "Machine learning in the electrocardiogram". In: *Journal of electrocardiology* 57 (2019), S61–S64.

[15] George B Moody and Roger G Mark. "The impact of the MIT-BIH arrhythmia database". In: *IEEE engineering in medicine and biology magazine* 20.3 (2001), pp. 45–50.

[16] Swathi Nayak et al. "Study of distance metrics on k-nearest neighbor algorithm for star categorization". In: *Journal of Physics: Conference Series*. Vol. 2161. IOP Publishing. 2022, p. 012004.

[17] Todd G Nick and Kathleen M Campbell. "Logistic regression". In: *Topics in biostatistics* (2007), pp. 273–301.

[18] F. Pedregosa et al. *Metrics and scoring: quantifying the quality of predictions*. 2023. URL: `https://scikit-learn.org/stable/modules/model_evaluation.html`.

[19] F. Pedregosa et al. *Scikit-learn: Machine Learning in Python*. Accessed: 2025-01-20. 2011. URL: `https://scikit-learn.org/stable/modules/neural_networks_supervised.html#neural-networks-supervised`.

[20] Siby Jose Plathottam et al. "A review of artificial intelligence applications in manufacturing operations". In: *Journal of Advanced Manufacturing and Processing* 5.3 (2023), e10159.

[21] Maryam Ramezani et al. "The application of artificial intelligence in health financing: a scoping review". In: *Cost Effectiveness and Resource Allocation* 21.1 (2023), p. 83.

[22] NL Rane et al. "Applications of machine learning in healthcare, finance, agriculture, retail, manufacturing, energy, and transportation: A review". In: *Applied Machine Learning and Deep Learning: Architectures and Techniques (112-131). Deep Science Publishing. https://doi. org/10.70593/978-81-981271-4-3_6* (2024).

[23] Steven J Rigatti. "Random forest". In: *Journal of Insurance Medicine* 47.1 (2017), pp. 31–39.