*Article*

# Deep Learning for Predicting Congestive Heart Failure

Francesco Goretti [1], Busola Oronti [2], Massimo Milli [3] and Ernesto Iadanza [4,*]

1    LENS, European Laboratory for Nonlinear Spectroscopy, University of Florence, Via Nello Carrara 1, 50019 Firenze, Italy
2    Applied Biomedical Signal Processing and Intelligent eHealth (ABSPIE) Lab, School of Engineering, University of Warwick, Coventry CV4 7AL, UK
3    USL Toscana Centro, Department of Cardiology, Ospedale S. Maria Nuova, 50122 Florence, Italy
4    Department of Medical Biotechnologies, University of Siena, Via Aldo Moro, 2, 53100 Siena Tuscany, Italy
*    Correspondence: ernesto.iadanza@unisi.it

**Abstract:** Congestive heart failure (CHF) is one of the most debilitating cardiac disorders. It is a costly disease in terms of both lives and financial outlays, given the high rate of hospital re-admissions and mortality. Heart failure (HF) is notoriously difficult to identify on time, and is frequently accompanied by additional comorbidities that further complicate diagnosis. Many decision support systems (DSS) have been developed to facilitate diagnosis and to raise the standard of screening and monitoring operations, even for non-expert staff. This is confirmed in the literature by records of highly performing diagnosis-aid systems, which are unfortunately not very relevant to expert cardiologists. In order to assist cardiologists in predicting the trajectory of HF, we propose a deep learning-based system which predicts severity of disease progression by employing medical patient history. We tested the accuracy of four models on a labeled dataset, composed of 1037 records, to predict CHF severity and progression, achieving results comparable to studies based on much larger datasets, none of which used longitudinal multi-class prediction. The main contribution of this work is that it demonstrates that a fairly complicated approach can achieve good results on a medium size dataset, providing a reasonably accurate means of determining the evolution of CHF well in advance. This potentially constitutes a significant aid for healthcare managers and expert cardiologists in designing different therapies for medication, healthy lifestyle changes and quality of life (QoL) management, while also promoting allocation of resources with an evidence-based approach.

**Keywords:** heart failure; deep learning; longitudinal data; electronic health records; trajectory

## 1. Introduction

Congestive Heart Failure, also known as Heart Failure, is a medical condition marked by decreased heart function and insufficient pumping action, resulting in an insufficient supply of oxygen and nutrients to meet the body's needs [1]. The main consequence of HF is low quality of life and increased treatment costs for sufferers over time. In particular, HF is accompanied by fatigue and short breath, leading to difficulties in normal day-to-day activities. Early detection of HF is critical for determining its causes and developing an effective treatment plan, which may include pharmacological or surgical options. An estimated number of 64.3 million people worldwide is currently living with a form of HF [2]. The death rate from HF is reported to be reducing in recent years [3], therefore encouraging further research into improving patient care. Nevertheless, HF still remains a serious condition with no definitive cure and should be treated as soon as possible to mitigate its progression. According to the 2021 European Society of Cardiology (ESC) guide [4], HF prevalence varies based on the definition given. It is diagnosed in approximately 1–2% of the adult population in developed countries, but could be higher due to unrecognized cases, and is mostly present in elderly people, especially those above 70 years. The incidence in developed countries is decreasing, probably because of better management of HF. However, because the average age is increasing, the overall incidence is increasing as well.

The etiology of HF depends on the geographical area being considered, as there is no single, general classification. The difficulty in assigning a proper definition comes from the fact that HF can manifest at a chronic stage of multiple possible overlapping diseases. Despite the challenge in identifying HF due to the presence of multiple comorbidities, it is fundamental to have a clear and complete clinical picture in order to design a good treatment plan. Moreover, it is difficult to give a prognosis on the exact evolution of HF and, as reported by Groenewegen et al. [2], physicians are often reluctant to communicate quantitative information about patients' health progression. Furthermore, the authors report how HF is still a big problem worldwide despite the evolution of therapies and clinical trials—a situation caused mainly by the increase in the average age of the elderly population. This implies that there are more people at higher risk, as well as a greater diffusion of HF in lower age groups. Notably, due to the high variability of HF cases, it is not easy to perform a correct diagnosis, specifically, in defining the quantitative evaluation of the failure. Unfortunately, HF treatment weighs heavily on both economic and medical resources globally. There has been an estimated doubling of cases from 1990 to 2017, with an increase of severe consequences, such as years lived with disability after HF, especially in nations with low socio-demographic status [5].

Cook et al. [6] analyzed 197 countries and obtained a worldwide evaluation of the economic burden of HF; it was estimated that 108 billion dollars have been spent to treat this condition, out of which 65 billion was used for direct costs of hospitalization, while the remaining part is correlated with lost productivity, morbidity and mortality. Recent studies prove that the situation does not appear to be improving; for example, in the United States, seven million new patients were discovered in 2020 and the approximate annual cost per patient was calculated to be 24,383 dollars [7]. Lesyuk et al. [8] evaluated a larger group of nations reporting an annual variable cost per patient, spanning from 868 dollars in Korea, to 25,532 in Germany, and 15,952 dollars spent in Italy. Generally, treatment costs increase as a result of various factors, most notably the severity of signs and symptoms [9], providing yet another incentive to improve the quality of diagnosis systems. The rise of HF patients, alongside high readmission rates of more than 30% in Europe and the United States [10], appear to make this disease one of the most serious concerns faced by healthcare systems. A recent review summarizes available information [11]. Reported numbers widely vary depending on the specific nation under observation, and in many countries incidence is declining. Unfortunately, on a worldwide scale, HF still represents one of the most burdensome cardiovascular diseases. Major reported statistics for HF include:

- High prevalence (1–3% in adult population).
- An incidence of 1–20 cases per 1000 population.
- High mortality: from 30 days mortality of 2–3% to five years mortality of 50–75%.
- Increased treatment costs. Mainly due to the increase of people over 65 years of age.

According to the studies referenced above, large sums of money are spent annually on hospitalization and treatment of HF patients. As well, lasting effects on lifestyle quality negatively impact the productive system, thus increasing the global burden of HF. All these issues reflect the need for efficient systems capable of slowing down the fast progression of the disease, as well as reducing hospitalization and indirect costs caused by HF. To overcome this challenge, numerous clinical decision support systems (CDSS) have been developed in the past decade; from simple systems based on rules, to modern algorithms based on Machine Learning (ML) and Deep Learning (DL).

ML is a subset of artificial intelligence (AI) based on algorithms capable of learning decision patterns directly from data [12]. What distinguishes ML from other types of AI is the ability to extract rules and discrimination boundaries from examples without the need of human intervention. ML can be further subdivided based on the type of data used in the learning phase (i.e., supervised if the examples are labeled, unsupervised if otherwise), or the expected output type (i.e., classification if there is a limited number of possible outputs, regression if the objective can assume continuous values).

DL [13] involves a group of learning algorithms based on artificial neural networks (ANNs) [14]. ANNs were inspired by the human brain structure. They consist of computing units called neurons with specific activation functions and parameters. The activation function integrates incoming signals in complex ways, permitting the resolution of very complicated problems. The main difference between DL and the neural networks used in ML is that in the case of DL, the layers are used to achieve space transformations and different representations of input data, whereas in classic ML, the layers just act as weight functions to filter the inputs (e.g., the perceptron). Also, when it comes to very large datasets or complex tasks, DL can transform data naturally within its structure without resorting to costly pre-processing stages. A typical representation of a simple perceptron is shown in Figure 1.
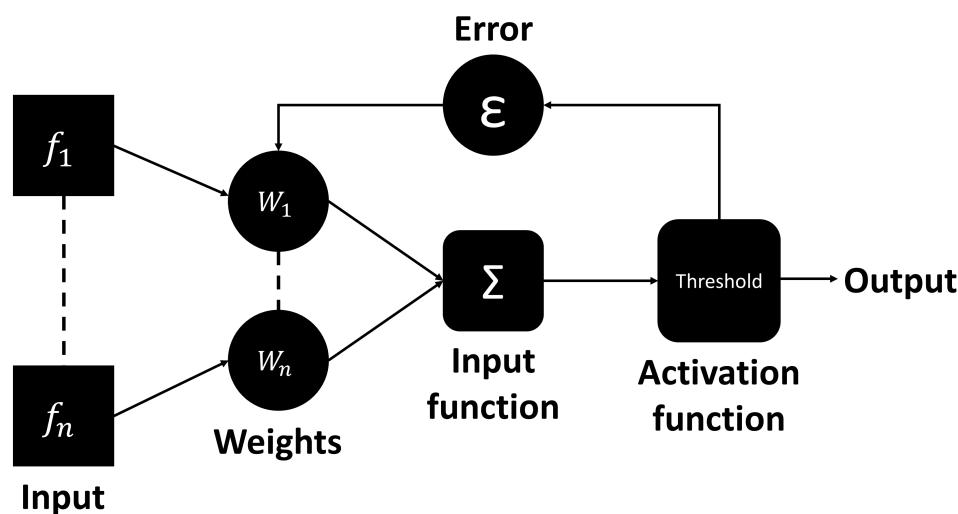
**Figure 1.** The perceptron rule. Graphical representation of the process of weight adjustment and the operation chain from the input to the classification.

Starting from year 2010, DL experienced an explosive evolution in terms of adaptation. This evolution specifically concerns the development of the following: (1) different layers and algorithms to tackle increasingly difficult tasks, and (2) networks for specialized functions. The most used DL networks are mainly Convolutional Neural Networks (CNN), typically used when images are involved, and Recurrent Neural Networks (RNN), which are capable of discriminating patterns, not only by the value of the features, but also from the order in which they are presented. Thus, they are particularly efficient for the study of longitudinal data. In this paper, we present the development and testing of a CDSS using DL techniques for HF severity prediction of the future entity (disease state) of HF patients. Our main goal is to aid expert clinical personnel in designing proper therapy for HF patients, thereby easing their disease burden. We used a dataset composed of patients at various stages of HF development (multiclass samples), and tested the performance of various DL models. The combination of these modern deep learning techniques with multi-class future entities as targets brings novelty to the current state of the art. To the best of our knowledge, no previous work has been done in this area. Moreover, gaining a foreknowledge of a patient's future HF state could provide an important treatment aid for domain experts, while also serving as a tool for managing clinical investments, personnel, and structures when dealing with chronic diseases. We anticipate that this system could also ultimately serve as an exemplar for testing new DL prediction algorithms for different pandemic-like diseases.

## 2. Related Work

DL is currently being used to analyse clinical data in many different contexts. However, use in medical imaging is perhaps the most ubiquitous and has witnessed the implemen-

tation of very advanced algorithms. Many new representative works in this field include brain Magnetic Resonance (MR) image super-resolution [15], optimized models with specific layers for the stratification of patients with cognitive disorders [16] and assessment of skeletal disorders in humans using X-ray images [17].

Of particular interest in this study is the evaluation of mortality and hospitalization due to HF since it is one of the greatest challenges associated with the disease. In a recent systematic review [18], many ML-based studies were reported to have dwelt on the prediction of these outcomes. The measured Area Under the Receiver Operating Characteristic (AUROC or AUC) curve varied from between 0.47 to 0.84 for hospitalizations, and from 0.48 to 0.92 for mortality predictions. Less accurate results frequently resulted from data scarcity and/or poor quality of data. Given the high variability in the symptoms of HF, the capability of ML algorithms to predict different objective functions has been tested; these include discrimination of HF patients and HF category, and the prediction of possible outcomes. Different types of data (electronic health records (EHR), electrocardiogram (ECG), etc.) have been employed in several studies to obtain these evaluations. Summarized results prove that ML can potentially outperform conventional methods, and will likely become the future technology for managing HF patients [19,20].

The area of binary classification for HF is well explored in the literature [21,22]. Various strategies and datasets have been tested with results reaching values up to 100% in accuracy, using different kinds of data. Vectors realized with EHR represent one of the simplest solutions to obtain an instance from each patient [23]. Guidi et al. [24] compared the accuracy of diagnosis of four different algorithms. They also proposed a portable interface for enhanced home monitoring, thus lowering the burden of patient hospitalizations and frequent hospital visits. The outcome of their work led to the realization of the dataset used in this study. Further work by Guidi et al. [25] incorporated more modern algorithms, which led to an improved dataset; specifically, a random forest-based system reached 89% accuracy in the classification of HF severity defined by clinicians. The management system was optimized in 2014 by the same authors [26] by realizing a complete interface and a database to manage HF patients. This made it possible to store data and visualize patients' progress, data and follow-up. Additionally, the system was implemented with a trained ML-based classifier to obtain an evaluation of the patient severity and the probability of re-hospitalization based on the considered instant.

The concept of prediction using longitudinal data is of great importance in the medical field. There are works in the literature that exploit ML and DL for prediction tasks even outside the field of HF [27,28].

Many different ML algorithms are easily available and offer the possibility to test different strategies to classify HF; Plati et al. [29] tested seven different ML algorithms in a dataset composed of EHRs from 422 subjects obtaining 92.23% accuracy with a rotation tree (ROT) that proved to be the best model.

Similarly, discriminating features can be extracted from Heart Rate Variability (HRV) measures [30]. Melillo et al. [31] were able to identify the features that best separate different NYHA classes from HRV analysis of a public dataset, after which they implemented a decision tree classifier, which achieved 95% accuracy.

The versatility of DL allows for improvement in the kind of data used for classification, while also allowing for the elimination of the features extraction phase. Furthermore, the performance of DL algorithms applied to HF classification has been shown to be reliable and efficient. Nirschl et al. [32] employed a CNN to detect HF from biopsy images and obtained 99% sensitivity and 94% specificity, using 105 and 109 images as training and test sets respectively. One of the main limitations of DL algorithms is the difficulty of obtaining explainable representation of the decision pattern of the model. Rao et al. [33] exploited the data of more than one hundred thousand patients to test the accuracy of a deep NN in the task of predicting the risk of developing HF in a six month window, obtaining 93% accuracy and an area under precision recall curve (AUPRC) of 0.69. They were also able to show the most important features, according to the algorithm, by

applying a process of features ablation. Likewise, sound signals have been tested with 92.5% classification accuracy achieved using a mixed ML/DL system [34]. Other authors have also examined audio recordings, and they were able to classify patients into various NYHA classes using recorded speech with 95% accuracy in a small sample of 31 patients [35]. Other unconventional data has been tested to classify HF, D'addio et al. [36] fed features, extracted from 199 poincarè plots, to AdaBoost, KNN and NB to classify patients based on the NYHA class reaching an accuracy of 80% and an AUROC of 0.7.

Finally, different studies have worked on extracting information from ECG signals as these provide a huge source of data. Kwon et al. [37] compared ML and DL algorithms to differentiate various types of HF based on the ejection fraction. Using a dataset composed of 55,163 ECGs, they observed that the DL algorithms outperformed the ML methods, with AUC greater than 0.820. In another study, single heartbeats were analyzed and the results obtained from five minute periods were combined to discriminate between healthy and HF subjects, reaching almost 100% accuracy [38]. DL allows the testing of many different structures and methods; Li et al [39] exploited a multi-scale residual network to classify 764 patients into different NYHA classes with an accuracy of 94%.

To the best of our knowledge, there are only two publicly available datasets similar to the ones used in this study: (1) the Heart Failure Prediction Data Set [40] and (2) the UCI Heart Disease Data [41]. Although they both deal with important aspects of heart diseases, mortality for the former and HF development for the latter, both differ greatly from our dataset in terms of the quantity of cases and the objective class. Numerous articles have tested and refined the categorization of these datasets' objective labels using ML algorithms, both in the fields of mortality prediction [42–44] and HF development [45–47].

## 3. Materials and Methods

### 3.1. Dataset

From 2012 to 2021, cardiology specialists at the Santa Maria Nuova Hospital in Florence, Italy, gathered a total of 1888 records from 760 patients. Several patient characteristics (grouped into five categories) were measured and recorded during each visit, namely, parameters, etiology, comorbidities, therapy and extracted values from ECG. The number of follow-up visits for each patient ranged from one to ten. For algorithmic necessity, these patient features were further sorted into three categories based on their numerical type. Table 1 shows the resulting classification of patient characteristics.

For algorithmic reasons, the features presented in Table 1 have been categorized by numeric type. Some of the most commonly measured continuous cardiac parameters include: (a) Systolic arterial pressure, (b) Diastolic arterial pressure (c) Cardiac frequency, and (d) Ejection fraction (EF). Normal values for systolic (maximum) and diastolic (minimum) pressure are usually 120 mmHg and 80 mmHg, respectively. The cardiac frequency is the number of complete cardiac cycles executed each minute; 60 pulses per minute (PPM) at rest is considered as normal frequency. The EF is the percentage of blood ejected by a specific chamber, which can either be the entire heart, the atrium, or ventricle. It is a useful index for measuring the contractile power of the heart and is often used to define different types of HF. The EF is calculated, as shown in Equation (1), as the difference between the end diastolic volume (EDV) and the end systolic volume (ESV), normalized with respect to the ESV, and multiplied by 100 to obtain a percentage value.

$$EF(\%) = \frac{EDV - ESV}{EDV} \times 100 \tag{1}$$

**Table 1.** Features recorded for each visit by physicians.

| Continuous | Ordinal | Boolean |
|---|---|---|
| Age | NYHA class | Ischemic heart disease |
| Systolic arterial pressure | ACE-blockers dose level | Hypertension |
| Diastolic arterial pressure | Sartans dose level | Valvulopathy |
| Weight | Beta-blockers dose level | Cardiomyopathy |
| Height | Diuretics dose level | Toxic heart disease |
| Cardiac frequency | compliance | Diabetes |
| Ejection fraction | | COPD |
| ACE-blockers Dose | | Kidney failure |
| Sartans dose | | Dyslipidemia |
| Betablockers dose | | Cerebrovascular pathologies |
| Diuretics dose | | Thyropathy |
| BNP (or proBNP) | | Hepatopathy |
| Oxigen saturation | | Sinus rhythm |
| | | Atrial fibrillation |
| | | Brachial block |
| | | Pacemaker ICD |
| | | Pacemaker ICDCRT |
| | | Digitalic |
| | | Antialdosterone |
| | | Antiplatelet agents |
| | | Anticoagulants |
| | | Nitrates |
| | | Statins |
| | | Amiodarone |
| | | Ivabradine |
| | | Surgical therapy |

BNP = Brain Natriuretic Peptide, NYHA = New York Heart Association, COPD = Chronic Obstructive Pulmonary Disease , ICD = Implantable Cardioverter Defibrillator, ICDCRT = Implantable Cardioverter Defibrillator Cardiac Resynchronization Therapy.

EDV is the remaining blood volume (with respect to the ventricle) at the end of the diastolic phase, while ESV is calculated at the end of the systolic phase. Also of particular note is the brain natriuretic peptide (BNP) [48]. BNP is part of the natriuretic peptides family, a group of substances produced by the heart. Typically, small amounts of BNP are found in the blood, and high levels are correlated to the insufficient capability of the heart in pumping blood. Due to the significance of BNP, a team of authors has created a quick colorimetric diagnostic kit based on an interaction between NT-ProBNP and a specific antigen [49].

Similarly, for ordinal values, the NYHA classification is a parameter that helps physicians in categorizing patients affected by HF. There are four different levels based on patient complexity. Finally, the Boolean features report a particular condition or specific medication. The number of valid records (with or without missing values) was 1739. The first data preprocessing task was to deal with missing/incorrect data, particularly for the BNP feature. Two methods were applied: firstly, with the help of our clinical partners, we recovered some missing BNP values for affected patients by assigning values acquired within two months from the time of data analysis; secondly, some incorrectly assigned values were removed upon critical examination of the remaining records, leaving a total of 641 records with a BNP or proBNP value assigned. For those cases where only proBNP values were available, we applied a change factor of four, as suggested by our clinical partners after discovering from a literature analysis [50–52] that conversion factors and formulas may vary as a function of the data source, nationality, and some other environmental issues.

Another challenge that emerged from the data analysis process was the difference in the number of patient follow-ups, for example, some patients had only four visits, while

some others had up to ten. To solve this problem, we adopted the padding strategy, which generally provides uniformity of dimensions. We shall describe this procedure in more detail in later sections. It is worthy of note, however, that some records could not be transformed into the desired dimension—a situation which translated into less knowledge of the patient's history. At the end of every cardiology visit, every patient was assigned a score representing the feature object of prediction (i.e., 1-mild, 2-moderate, and 3-severe), which gives an indication of the severity of the HF. Finally, a preliminary data-cleaning phase was executed by removing records with too many missing values (>3) and features with low variance (<0.8). All involved patients signed the consent to process their personal data. It is useful to emphasize here that we received already anonymized data from our clinical partners. Moreover, all data records had been collected following the usual standard of care procedure. No additional patient examination was executed for the purpose of this study.

### 3.2. Prediction System Development

To build the prediction system for HF, we developed a multi-input neural network using DL techniques. ML algorithms were not tested because of the heterogeneous nature of the follow-up counts, since it was not possible to establish a criterion to build a dataset valid for ML, which also carried temporal information. A common approach, applied by Zhao et al. [53], is to line up different follow-up counts. This method requires a large quantity of longitudinal data registered consistently over many years without missing values. Unfortunately, the nature of data we were dealing with could not accommodate these requirements.

As the main goal of this study is to predict the evolution of HF entity by designing a multi-class prediction system based on longitudinal EHR data, we adapted the methods presented by Zhao et al. [53] and Rongali et al. [54], respectively. In the former, the authors applied ML and DL models to EHR longitudinal and genetic data with the goal of predicting a ten-year cardiovascular event outcome for individuals, while in the latter, this concept was extended by developing a model capable of identifying uniform representations between different types of clinical features for each patient encounter. Following this, an RNN stage was developed to add temporal information to the data, which was then split into 3 sets, i.e., training and validation sets for model selection, and the test set for final evaluation. In our adaptation of the combination of these two methods for deploying multi-input neural networks, we had the opportunity to explore an uncommon path with a unique dataset of HF patients, each labelled by expert clinical personnel in nine years of hospital visits. All the executed processes, starting from the data pre-processing to the test phase, are presented in Figure 2 and have been applied to each network structure separately, thus making it possible to evaluate the best model for each structure on the test set.

### 3.2.1. Data Pre-Processing

The data pre-processing stage was designed to produce the best data conditioning specified by DL theory. The procedure for data pre-processing can be divided into four stages:

- BNP interpolation
- Cleaning
- Scaling
- Structuring data

As mentioned earlier, missing BNP values in our dataset posed a major challenge. To solve this problem, we set a threshold value (criterion) for each of the different classes of HF as follows:

- Class Mild: BNP < 500
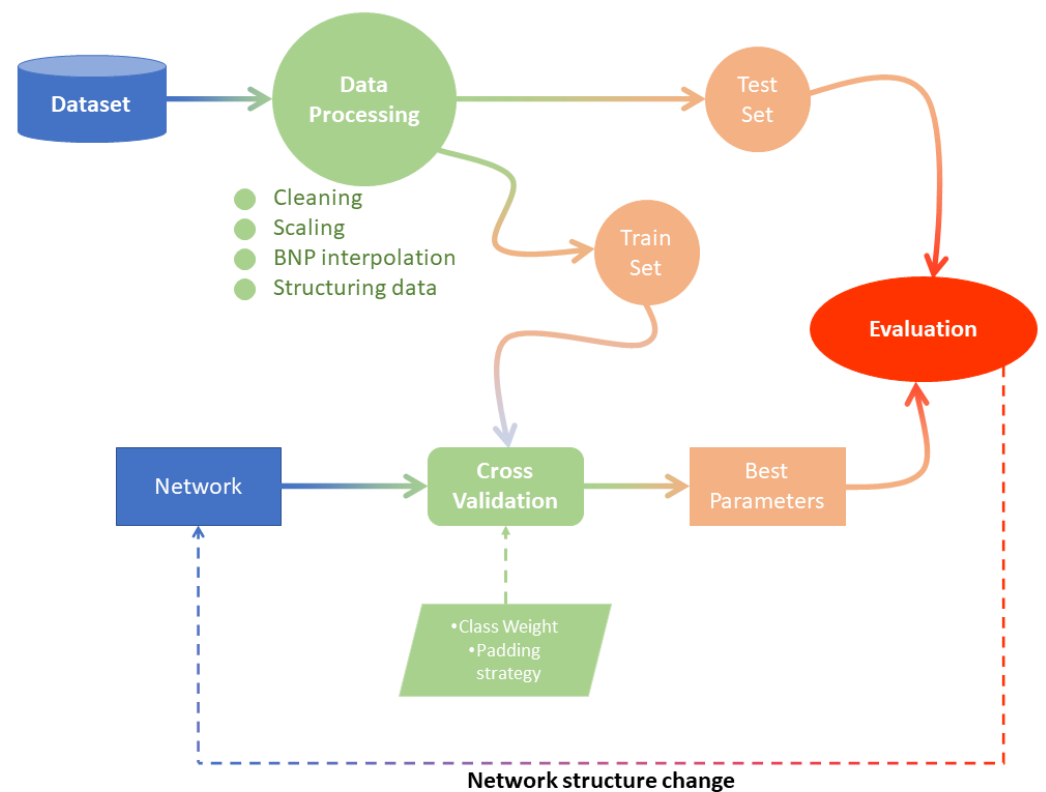- Class Moderate: 500 < BNP < 1000
- Class Severe: BNP > 1000

**Figure 2.** Flow chart of developed model.

Each criterion was applied by generating a random number in the selected range based on the class of the record with missing BNP values. We were confident that this approach would not bias our data since it was applied by looking at the actual class, while keeping the objective label as the future class. Once the dataset was repopulated, patient records with a single visit were deleted. This is because for prediction tasks, the first entity value of the series of visits is unusable due to lack of additional data to implement the task. Next, we deleted those features which presented low variance (<0.8), and would have provided minimal information to our model. Unfortunately, many patient records had a lot of missing values and the threshold adopted to remove these data instances also deleted records with more than three unrecoverable features. Consequently, the number of complete records for our study was composed of 1037 instances with complete features (after the BNP interpolation). Beyond these, no further features selection was performed since the number of instances obtained was large enough to implement the model.

For the second step of this stage, standard scaling was applied to the ordinal and numerical data respectively, as neural networks have been shown to work better with scaled data. The scaling method employed (standard scaling) transformed the value of each feature so that they had mean equal to 0 and variance equal to 1. This method is effected via the following equation:

$$Sv = \frac{x - \mu}{\sigma} \qquad (2)$$

where $Sv$ is the scaled value, $x$ the original feature value, $\mu$ the average of the feature values and $\sigma$ the standard deviation.

Scaling was not applied to the Boolean group because these features can only assume a value of 0 or 1, and scaling would be useless. Thereafter, the whole dataset was divided into the training and test sets in the ratio 90/10. This division was made with respect to two important considerations; firstly, that the training and test sets must have the same class distribution, and secondly, that data coming from the same patient must either be in the training or test set. Furthermore, we implemented a function for the creation of a triple

input dataset and encoded the objective labels with a one hot strategy. The purpose of incorporating this step was to recreate the progressive knowledge that physicians gain with successive patient visits and use the information acquired to predict the future entity of the pathological condition. To exploit all possible information from our dataset, we generated from each series of N visits a number of instances equal to N-1. This provided growing knowledge owing to the increment in patient history information as shown in Figure 3. Since neural networks accept inputs of the same dimensions, we applied the padding strategy to unify the size of temporal data. This is basically a method of filling missing dimensions with constant values. Two different padding strategies were tested, namely, zero padding (fills all missing dimensions with series of zeros) and custom padding (fills all missing dimensions with the oldest value).

The function for implementing the data instances, which was used to train and test the model, makes use of the original list of features, and returns three elements for each patient. The first element is a 2D array composed of stacked temporal values, while the second and third are arrays with Boolean and ordinal values, respectively. The last column displays the HF entity value of the next (future) visit for each patient. For this reason, the first label of the series and the last array of values are excluded from the prediction system. Finally, the 2D arrays were made uniform in shape by the application of padding operations as described previously.

### 3.2.2. Models Design

According to the literature, the best results in the field of HF prediction or classification, using EHR longitudinal data, have been achieved with RNN-based networks [53–56]. It is well known that there is no general, golden rule for constructing ANNs [57,58], and there are no specific networks which work perfectly for certain tasks (except for some considerations on the algorithms used). Given this scenario, we started with a simple baseline model and repeatedly checked variations in performance as the parameters changed. Specifically, the temporal branch is responsible for managing features which define a temporal dependence and is composed of RNN layers, while the ordinal and Boolean branches extract features from simpler data with densely connected layers. Patient features were therefore reordered and grouped for the multi-input neural network as follows:

- Temporal data: Weight, ejection fraction, BNP, NYHA class, age, systolic arterial pressure, diastolic arterial pressure, cardiac frequency, ACEblockers dose, Betablockers dose, diuretics dose.
- Boolean data: Ischemic heart disease, hypertension, dyslipidemia, sinus rhythm, atrial fibrillation, diabetes, BPCO, nitrates.
- Ordinal data: Sarta dose level, Betablockers dose level, compliance.

There are many different methods to encode variables and features in order to properly feed them to an ANN [59]; amongst them continuous data were scaled to have mean equal to 0 and unit variance, ordinal encoding was chosen for the ordinal features and boolean features were already in the binary form. Finally, labels were encoded according to the one-hot procedure. Given the relatively small shape of the dataset, standard encoding operations were applied to the features. As can be seen above, the NYHA class was reordered from ordinal to temporal data since the variation of this feature is important in the evaluation of the HF entity. Finally, a concatenated layer was created to merge all representations obtained from different data types to give the final prediction. A graphical representation of the network design is shown in Figure 4.

We tested a total of four different models. The first model was taken as base, and then different permutations were tested. The second model was realized to see if adding complexity would bring some improvement in performance, while for the third and fourth models, we implemented the network structures designed by Zhao et al. [53]. Two aspects of the third and fourth models are worthy of note: firstly, due to differences in the strategy adopted for data organization, we copied the structure for the temporal branch only, and changed the last layer to discriminate between three classes instead of two; secondly, for

the fourth model, we applied a different approach using a 1D CNN instead of RNN layers. All the models used dropout layers [60], which work by randomly dropping basic neurons (along with their connections) from the neural network during training to overcome the issue of overfitting.

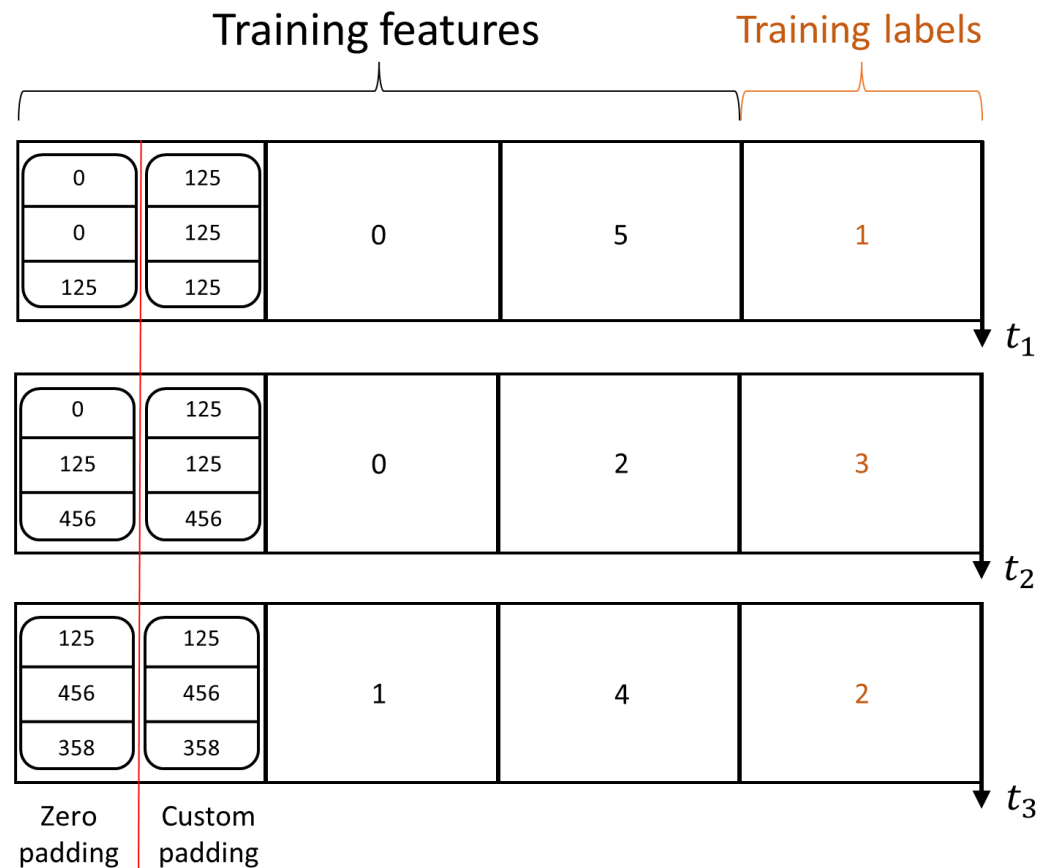| Time dependant | Boolean | Ordinal | Severity |
|---|---|---|---|
| 125 | 0 | 5 | 2 |
| 456 | 0 | 2 | 1 |
| 358 | 1 | 4 | 3 |
| 286 | 1 | 4 | 2 |



**Figure 3.** Visual representation of the dataset creation process. An example of data instances obtained from a single patient history. From a history of length N, N-1 data instances can be created since the most recent record cannot be associated with a future severity. Labels are represented in red.

The main difference between the models was the type and number of layers. The temporal branches were composed of long short term memory (LSTM) layers, while the other classes of data were processed by dense layers. The maximum depth of the models varied from five to ten layers, each with 32 or 64 inner units. The fourth model represents an exception since the LSTMs were replaced with a simple CNN structure. The optimizer chosen for the compilation was based on the Adam algorithm [61] with a learning rate of 0.001. The loss metric we chose to minimize is the categorical cross entropy, since this is suggested when dealing with multi-class problems [58]. Regarding training parameters, the batch size was fixed as 32 and the epochs selected based on the early stopping method.

For each model, we applied 10 fold cross-validation (CV) and calculated the required performance metrics at each instance after changing three sets of weights and two padding strategies, which resulted in a total of six assessments for each model. The CV was used to define the best sub-model by changing the weights and the padding method.
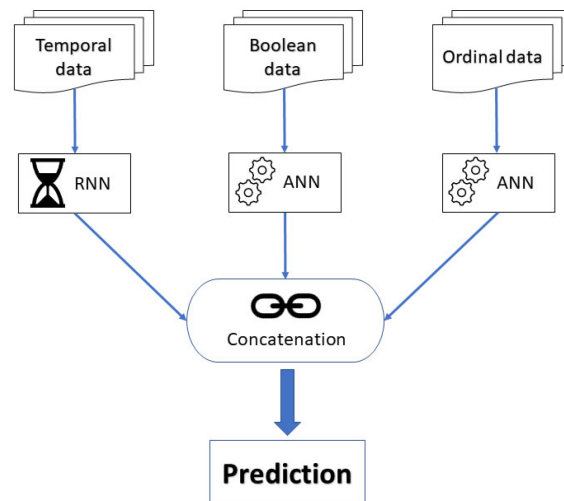


**Figure 4.** Sketch of the designed multi input neural network. Different types of data are processed by specific layers. Information from data that may have temporal evolution are extracted by a RNN, while other data are passed through dense layers. The last concatenated layer merges the features extracted by all the branches to provide the final prediction.

## 4. Results

The metrics considered are as follows:

- True Positives (TP): Positive instances correctly classified
- False Positives (FP): Negative instances classified as positive
- True Negatives (TN): Negative instances correctly classified
- False Negatives (FN): Positive instances classified as negative
- Accuracy: Represents the number of correct predictions with respect to the total number of samples
$$Acc = \frac{TP + TN}{TP + TN + FP + FN}.$$
- Recall or Sensitivity: Ratio of positive instances correctly classified to total (actual) positives in the dataset $Rec = \frac{TP}{TP + FN}$.
- Precision: Accuracy of the positive predictions $Prec = \frac{TP}{TP + FP}$.
- False positive rate (FPR): Ratio of false positives to the total number of actual negative events $\frac{FP}{FP + TN}$
- Area Under Receiver Operating Characteristic (AUROC) curve: Area under the Receiver Operator Characteristic (ROC) curve, which plots the true positive rate (recall) against the FPR. To plot the entire curve, these two metrics are evaluated many times after the variation of a classification threshold. The best value for the AUROC is 1.
- Area Under Precision Recall Curve (AUPRC): This metric is particularly used for binary responses. It is appropriate for rare events and is not dependent on model specificity. In this case, the axes are defined as precision and recall, respectively.
- Confusion matrix: This matrix is the most complete way of representing results. It is shown as a table containing true values in the rows and predicted values in the columns. A perfect confusion matrix is diagonal; values in the diagonal are predicted correctly and the others are not.

For multi-class classification, the usual approach is the One Versus All (OVA), which calculates relevant metrics by iteratively defining one of the classes as positive, and the others as negative. The final value is the mean of all the partial results.

The set of parameters which provided the best results was passed through a final evaluation with the test set to check its generalization capabilities. This final test was executed with a number of training epochs equal to the mean of the total epochs obtained during the 10 experiments with the training set. This last detail was considered due to the fact that we used early stopping during the CV to obtain the best value before overfitting (associated with too many training epochs) could take effect. The padding strategies adopted have been discussed in the previous section. To address the issue of class imbalance, assigning a weight representative of the original class distribution is a common approach, and is recommended when using Python algorithms to feed scaled values. The reason for this is because high values can result in poor model performance and flip the bias on the classes. To counter this effect, we implemented a function to logarithmically scale the count of the classes by multiplying with a variable number, which gave us three different sets of weights. The weight assigned to the majority class is always 1, i.e., the moderate HF class in our case.

Results without the use of weights were not reported in this paper since almost all the predictions were of the majority class as a result of the bias introduced by the unbalanced classes. In Table 2, we report the results of the 10 fold CV, while the confusion matrices, shown in Table 3, present the results on the test set of the model trained with the best set of parameters for each of the four models. The best model in terms of considered metrics is presented in bold text.

**Table 2.** Results of 10-fold CV of the best models. In bold the best model according to the selected metrics.

| Model | Accuracy | | | AUROC | | | AUPRC | | | Parameters |
|---|---|---|---|---|---|---|---|---|---|---|
| | mean | std | variance | mean | std | variance | mean | std | variance | |
| 1 | 61.71 | 4.65 | 51.58 | 79.15 | 3.45 | 11.93 | 59.67 | 6.13 | 37.63 | 0 padding Weight set 1 |
| 2 | 60.89 | 4.25 | 18.02 | 77.99 | 3.49 | 12.20 | 58.71 | 6.12 | 37.40 | 0 padding Weight set 1 |
| 3 | 61.56 | 4.36 | 18.98 | 78.64 | 2.50 | 6.25 | 57.84 | 4.46 | 19.86 | Custom padding Weight set 1 |
| **4** | **62.74** | **1.55** | **2.39** | **79.24** | **1.68** | **2.82** | **61.64** | **3.71** | **13.74** | **0 padding Weight set 1** |

**Table 3.** Confusion matrices of best models on the test set. In bold the model with the best trade-off between accuracy and no-bias toward the majority class.

| Model 1 | | Predicted Labels | | | Model 2 | | Predicted Labels | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mild | Moderate | Severe | | | Mild | Moderate | Severe |
| True | Mild | 11 | 12 | 0 | True | Mild | 11 | 11 | 1 |
| labels | Moderate | 3 | 53 | 12 | labels | Moderate | 5 | 45 | 18 |
| | Severe | 0 | 4 | 5 | | Severe | 0 | 2 | 7 |
| Model 3 | | Predicted Labels | | | Model 4 | | Predicted labels | | |
| | | Mild | Moderate | Severe | | | Mild | Moderate | Severe |
| True | Mild | 8 | 13 | 2 | True | Mild | 15 | 8 | 0 |
| labels | Moderate | 3 | 34 | 31 | labels | Moderate | 13 | 55 | 0 |
| | Severe | 0 | 1 | 8 | | Severe | 0 | 9 | 0 |

In the first stage of our evaluation, model number 4 appeared to be the best, with an accuracy of 62.7%, AUROC of 79.3 and an AUPRC of 61.7. The ROC curves, reported in

Figure 5 show that all the models perform better than a random classifier, but still achieve similar scores, even on the test set. However, by looking at the confusion matrices of the same set, it became clear that there is a bias toward the major class in model 4, leaving model 1 as the most balanced. For this reason, we decided to select the best performing model based on a compromise in performance on the training and test sets. Using this approach, model number 1 emerged as the best since it is the only model that does not output extremely wrong classifications, such as classifying the mild class as severe or vice versa. The confusion matrix is of particular benefit in presenting results because it can represent multi-class classification problems in a simple manner and is easy to understand. The final accuracy score on the test set is therefore 69%.
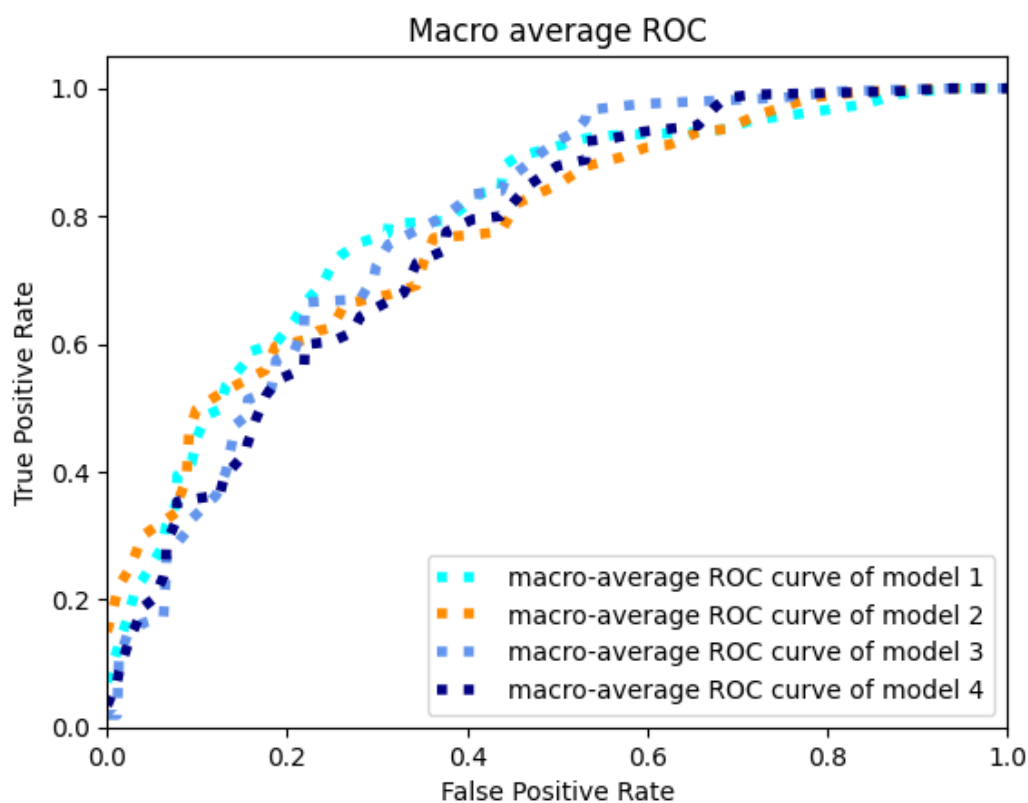


**Figure 5.** ROC curves on the test set. All the models performed similarly on unseen data. Examination of other metrics provides deeper insights.

## 5. Discussion

In this work, we realized a deep learning-based system capable of predicting the entity of HF on a custom dataset using longitudinal EHR. Varying some of the possible network parameters did not result in any significant changes in the model performance. This suggests that there might have been an inherent performance threshold we were not able to scale, probably due to some limitations in the dataset or perhaps because we did not manipulate the correct hyperparameters. Further experiments with a larger pool of data are therefore needed to ascertain this possibility. Within the scope of our research, we have not found any similar study involving the use of mixed-type, longitudinal, EHR data for multi-class system classification, where the objective is to predict the evolution of HF, nor studies that specifically used the same features as ours to implement their model, thus making it hard to compare our results with others obtained in the literature. ML for HF-related tasks has been widely used in many studies, using different kinds of data (e.g., ECG, EHR, etc.) to obtain indications related to the diagnosis or risk analysis [18,19]. A reduced version of the dataset used in this study was utilized in other studies as previously cited in the introductory section. The classification of two separate labels, namely, HF

severity and the type of HF, was examined and compared using neural networks (NN), support vector machine (SVM), fuzzy algorithm, decision tree (DT), and random forests (RF). In the cited work, 'type of HF' was adopted as a label given by doctors based on how many times the patient was readmitted to the hospital in a year. The best algorithms attained an accuracy of about 88% in the classification of HF type, and between 83% to 89% in the classification of severity [25,26]. In a later study, the same authors merged the system on a bigger framework for the management of HF patients, and re-tested the algorithms. Accuracy values approached 71.9% for HF type and 81.7% for HF severity [62]. The authors' best outcomes came from a different dataset composed of ECG records, where they were able to classify heartbeats as belonging to patients with or without HF with nearly 100% accuracy [38]. When dealing with longitudinal data, system complexities increase and results do not usually reach the performance levels of standard EHR-based classifiers. The usage of RNNs allows for extracting patterns based on the order in which data is presented to the model and represents one of the most novel approaches used when dealing with longitudinal records. Applications of such networks can involve a variety of HF tasks, ranging from event outcome prediction to re-hospitalization risk evaluation. One of the closest studies to our current work is probably the one implemented by Zhao et al. [53], where the authors applied ML and DL models to more than one hundred thousand EHR longitudinal and genetic data to predict 10-year cardiovascular event outcomes. In their work, they applied 8 fold CV and achieved values of AUROC between 0.78 and 0.80, and AUPRC between 0.25 and 0.27 with the best models. We obtained comparable results in terms of AUROC, and better results for the AUPRC values. Another comparable work is that of Rondali et al. [54] where they obtained a higher AUROC of 0.8. However, they implemented a binary mortality prediction and had access to a larger, and more homogeneous pool of data.

Other studies involving deep learning and longitudinal data are those of Golas et al. [55], where the authors obtained an AUC of 0.705 and accuracy of 64% on a 30 day readmission prediction task, as well as that of Choi et al. [56], where the authors obtained an AUC of 0.883 with their best model for the early detection of HF. Both works involved pools of data of approximately thirty thousand instances.

Adding the concept of the "trajectory" of HF disease led to the discovery of other similar works. However, there are several aspects that make these hard to compare with our work. For instance, Pham et al. [63] realized a DL-based system which used electronic medical records as input. They exploited the information obtained from the progression history of patients to provide two support outputs: one for the inference of the current illness state, and the other for the prediction of future outcomes. The pathologies under study were diabetes and mental health factors. The custom network they developed outperformed classic Markov models and standard RNN with different objectives to classify. Metrics scores were not incredibly high, but, considering the complexity of the task, the results highlight the importance of increasing studies in the field of predictive modelling. In another study, HF and disease trajectory were associated in a binary prediction task [64]. The authors proposed an RNN-based method to predict HF onset at the next hospital visit (as we did with our three HF classes). The goal was to build a trajectory sequentially predicting an event, and then use this as a base for further predictions. They used a dataset composed of more than 84 thousand patients and 28 years of follow-up. For the first task, they obtained an AUROC greater than 0.85, while the second task yielded an AUROC ranging between 0.8 and 0.6. Notably, AUROC values decreased when the trajectory to build became longer, specifically, from one to ten years. In some other studies, researchers have tried to predict other factors related to cardiovascular diseases. Guo et al. [65], for example, obtained very high values of AUROC for multiple predictions with a dataset of more than 200 thousand patients. The aim was to predict successive values of a series of cardiovascular health indexes such as smoking status or body mass index.

Summarizing, ML applications have been widely analyzed in the field of HF and standard techniques have been applied to different sets of data for diagnosis and risk

predictions. Longitudinal data is employed in a limited number of studies, and to the best of our knowledge, multi-class entity prediction has never been considered. With the help of our clinical partners, we have been able to use a unique longitudinal dataset obtained from HF patients during periodic hospital visits to explore this new area and propose a novel methodology for predicting HF entities. This system was designed to go beyond the diagnosis, which is to date a well investigated area, providing expert cardiologists with a novel tool for supporting the prognosis task. According to our clinical partners, this support system has the potential to become a game changer, useful to save time, money and to gain more confidence in communicating the prognosis to the patient. Many aspects represent obstacles in the present configuration, but gathering more data and the fast development of DL makes it reasonable to think that the performance will grow significantly in future studies The discordance in the number of follow-up visits amongst patients and missing values represent the main limitations of this work. We hope that the continued gathering of quality longitudinal data will help in designing a more solid protocol, capable of unifying all required dimensions. Despite these reported limitations however, our developed model has shown improved performance on the test set by gaining more accuracy and providing good generalization capabilities on data which had never been seen before by the algorithm. By exploring and applying hyperparameter tuning and advanced network optimization methods in future studies, we hope to discover the root cause of the performance limitations (whether algorithmic or data-related) highlighted in the present study and develop new methodologies and techniques that will aid better model performance.

## 6. Conclusions

The goal of this prediction system is to help expert cardiologists in the difficult task of predicting the evolution of HF, based on the clinical history of patients. This research output can help in adjusting or designing therapies that are best suited to individual patients, while taking into consideration HF unpredictability. In addition, the proposed approach can help organizations that manage health investments, particularly in relation to diseases that have high costs in terms of human and financial resources. Finally, future works may investigate the applicability of novel methodologies on this dataset, given the daily expansion of studies involving HF and ML. Fine tuning specific model parameters and employing more articulated networks could also be explored when more data become available. Likewise, a regression-type output could be deployed, where risk percentages for all classes could be tested in order to give more information to clinicians. It is hoped that in the very near future, greater opportunities will arise for sharing public datasets dedicated to thorough and state-of-the-art research in multi-class classification and/or prediction of HF or other chronic diseases, through the realization of more data repositories focused on this field of study.

**Abbreviations**

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AI | Artificial Intelligence |
| ANN | Artificial Neural network |
| AUPRC | Area Under Precision Recall Curve |
| AUROC or AUC | Area Under ROC |
| BNP | Brain Natriuretic Peptide |
| CDSS | Clinical Decision Support System |
| CHF | Congestive Heart Failure |
| CNN | Convolutional Neural network |
| COPD | Chronic Obstructive Pulmonary Disease |
| DL | Deep Learning |
| ECG | Electrocardiogram |
| EDV | End Diastolic Volume |
| EF | Ejection Fraction |
| EHR | Electronic Health Records |
| ESC | European Society of Cardiology |
| ESV | End Systolic Volume |
| FP | False Positive |
| FN | False Negative |
| HRV | Heart Rate Variability |
| HF | Heart Failure |
| ICD | Implantable Cardioverter Defibrillator |
| ICDCRT | Implantable Cardioverter Defibrillator Cardiac Resynchronization Therapy |
| KNN | K-Nearest Neighbors |
| LMT | Logistic Model Tree |
| ML | Machine Learning |
| NB | Naive Bayes |
| NN | Neural Network |
| NYHA | New York Heart Association |
| PPM | Pulse Per Minute |
| RF | Random Forest |
| RNN | Recurrent Neural network |
| ROC | Receiver Operating Characteristic |
| ROT | Rotation Tree |
| SVM | Support Vector machine |
| TP | True Positive |
| TN | True Negative |

**References**

1. Malik, A.; Brito, D.; Vaqar, S.; Chhabra, L. Congestive heart failure. In *StatPearls [Internet]*; StatPearls Publishing: Treasure Island, FL, USA, 2022.
2. Groenewegen, A.; Rutten, F.H.; Mosterd, A.; Hoes, A.W. Epidemiology of heart failure. *Eur. J. Heart Fail.* **2020**, *22*, 1342–1356. Available online: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejhf.1858 (accessed on 1 November 2022). [CrossRef] [PubMed]
3. Jones, N.R.; Roalfe, A.K.; Adoki, I.; Hobbs, F.R.; Taylor, C.J. Survival of patients with chronic heart failure in the community: A systematic review and meta-analysis. *Eur. J. Heart Fail.* **2019**, *21*, 1306–1325. [CrossRef]
4. McDonagh, T.A.; Metra, M.; Adamo, M.; Gardner, R.S.; Baumbach, A.; Böhm, M.; Burri, H.; Butler, J.; Čelutkienė, J.; Chioncel, O.; et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC) With the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur. Heart J.* **2021**, *42*, 3599–3726. [PubMed]
5. Bragazzi, N.L.; Zhong, W.; Shu, J.; Abu Much, A.; Lotan, D.; Grupper, A.; Younis, A.; Dai, H. Burden of heart failure and underlying causes in 195 countries and territories from 1990 to 2017. *Eur. J. Prev. Cardiol.* **2021**, *28*, 1682–1690. Available online: https://academic.oup.com/eurjpc/advance-article-pdf/doi/10.1093/eurjpc/zwaa147/36239071/zwaa147.pdf (accessed on 1 November 2022). [CrossRef]
6. Cook, C.; Cole, G.; Asaria, P.; Jabbour, R.; Francis, D.P. The annual global economic burden of heart failure. *Int. J. Cardiol.* **2014**, *171*, 368–376. [CrossRef] [PubMed]

7. Urbich, M.; Globe, G.; Pantiri, K.; Heisen, M.; Bennison, C.; Wirtz, H.S.; Di Tanna, G.L. A systematic review of medical costs associated with heart failure in the USA (2014–2020). *Pharmacoeconomics* **2020**, *38*, 1219–1236. [CrossRef] [PubMed]

8. Lesyuk, W.; Kriza, C.; Kolominsky-Rabas, P. Cost-of-illness studies in heart failure: A systematic review 2004–2016. *BMC Cardiovasc. Disord.* **2018**, *18*, 1–11. [CrossRef]

9. Shafie, A.A.; Tan, Y.P.; Ng, C.H. Systematic review of economic burden of heart failure. *Heart Fail. Rev.* **2018**, *23*, 131–145. [CrossRef]

10. Ambrosy, A.P.; Fonarow, G.C.; Butler, J.; Chioncel, O.; Greene, S.J.; Vaduganathan, M.; Nodari, S.; Lam, C.S.; Sato, N.; Shah, A.N.; et al. The Global Health and Economic Burden of hospitalizations for Heart Failure. *J. Am. Coll. Cardiol.* **2014**, *63*, 1123–1133. Available online: https://www.jacc.org/doi/pdf/10.1016/j.jacc.2013.11.053 (accessed on 1 November 2022). [CrossRef]

11. Savarese, G.; Becher, P.M.; Lund, L.H.; Seferovic, P.; Rosano, G.M.C.; Coats, A.J.S. Global burden of heart failure: A comprehensive and updated review of epidemiology. *Cardiovasc. Res.* **2022**, cvac013. Available online: https://academic.oup.com/cardiovascres/advance-article-pdf/doi/10.1093/cvr/cvac013/43972759/cvac013.pdf (accessed on 1 November 2022). [CrossRef]

12. El Naqa, I.; Murphy, M.J. What is machine learning? In *Machine Learning in Radiation Oncology*; Springer: Cham, Switzerland, 2015; pp. 3–11.

13. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]

14. Mcculloch, W.S.; Pitts, W. A logical calculus nervous activity. *Bull. Math. Biol.* **1990**, *52*, 99–115. [CrossRef] [PubMed]

15. You, S.; Lei, B.; Wang, S.; Chui, C.K.; Cheung, A.C.; Liu, Y.; Gan, M.; Wu, G.; Shen, Y. Fine perceptive gans for brain mr image super-resolution in wavelet domain. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**. *Online ahead of print.*

16. Yu, W.; Lei, B.; Ng, M.K.; Cheung, A.C.; Shen, Y.; Wang, S. Tensorizing GAN with high-order pooling for Alzheimer's disease assessment. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 4945–4959. [CrossRef]

17. Wang, S.; Wang, X.; Shen, Y.; He, B.; Zhao, X.; Cheung, P.W.H.; Cheung, J.P.Y.; Luk, K.D.K.; Hu, Y. An ensemble-based densely-connected deep learning system for assessment of skeletal maturity. *IEEE Trans. Syst. Man Cybern. Syst.* **2020**, *52*, 426–437. [CrossRef]

18. Mpanya, D.; Celik, T.; Klug, E.; Ntsinjana, H. Predicting mortality and hospitalization in heart failure using machine learning: A systematic literature review. *IJC Heart Vasc.* **2021**, *34*, 100773. [CrossRef] [PubMed]

19. Bazoukis, G.; Stavrakis, S.; Zhou, J.; Bollepalli, S.C.; Tse, G.; Zhang, Q.; Singh, J.P.; Armoundas, A.A. Machine learning versus conventional clinical methods in guiding management of heart failure patients—A systematic review. *Heart Fail. Rev.* **2021**, *26*, 23–34. [CrossRef]

20. Tripoliti, E.E.; Papadopoulos, T.G.; Karanasiou, G.S.; Naka, K.K.; Fotiadis, D.I. Heart failure: Diagnosis, severity estimation and prediction of adverse events through machine learning techniques. *Comput. Struct. Biotechnol. J.* **2017**, *15*, 26–47. [CrossRef]

21. Krittanawong, C.; Virk, H.U.H.; Bangalore, S.; Wang, Z.; Johnson, K.W.; Pinotti, R.; Zhang, H.J.; Kaplin, S.; Narasimhan, B.; Kitai, T.; et al. Machine learning prediction in cardiovascular diseases: A meta-analysis. *Sci. Rep.* **2020**, *10*, 16057. [CrossRef]

22. Olsen, C.R.; Mentz, R.J.; Anstrom, K.J.; Page, D.; Patel, P.A. Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure: Machine learning in heart failure. *Am. Heart J.* **2020**, *229*, 1–17. [CrossRef]

23. Samuel, O.W.; Asogbon, G.M.; Sangaiah, A.K.; Fang, P.; Li, G. An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. *Expert Syst. Appl.* **2017**, *68*, 163–172. [CrossRef]

24. Guidi, G.; Iadanza, E.; Pettenati, M.; Milli, M.; Pavone, F.; Biffi Gentili, G. Heart failure artificial intelligence-based computer aided diagnosis telecare system. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7251 LNCS, pp. 278–281. [CrossRef]

25. Guidi, G.; Pettenati, M.; Miniati, R.; Iadanza, E. Random forest for automatic assessment of heart failure severity in a telemonitoring scenario. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, Osaka, Japan, 3–7 July 2013; IEEE: New York, NY, USA, 2013; pp. 3230–3233. [CrossRef]

26. Guidi, G.; Pettenati, M.C.; Melillo, P.; Iadanza, E. A machine learning system to improve heart failure patient assistance. *IEEE J. Biomed. Health Inform.* **2014**, *18*, 1750–1756. [CrossRef] [PubMed]

27. Miao, L.; Guo, X.; Abbas, H.T.; Qaraqe, K.A.; Abbasi, Q.H. Using machine learning to predict the future development of disease. In Proceedings of the 2020 International Conference on UK-China Emerging Technologies (UCET), Glasgow, UK, 20–21 August 2020; pp. 1–4.

28. Schvetz, M.; Fuchs, L.; Novack, V.; Moskovitch, R. Outcomes prediction in longitudinal data: Study designs evaluation, use case in ICU acquired sepsis. *J. Biomed. Inform.* **2021**, *117*, 103734. [CrossRef] [PubMed]

29. Plati, D.K.; Tripoliti, E.E.; Bechlioulis, A.; Rammos, A.; Dimou, I.; Lakkas, L.; Watson, C.; McDonald, K.; Ledwidge, M.; Pharithi, R.; et al. A machine learning approach for chronic heart failure diagnosis. *Diagnostics* **2021**, *11*, 1863. [CrossRef] [PubMed]

30. Hussain, L.; Awan, I.A.; Aziz, W.; Saeed, S.; Ali, A.; Zeeshan, F.; Kwak, K.S. Detecting Congestive Heart Failure by Extracting Multimodal Features and Employing Machine Learning Techniques. *BioMed Res. Int.* **2020**, *2020*, 4281243. [CrossRef]

31. Melillo, P.; Pacifici, E.; Orrico, A.; Iadanza, E.; Pecchia, L. Heart rate variability for automatic assessment of congestive heart failure severity. In Proceedings of the XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013, Seville, Spain, 25–28 September 2013; Volume 41, pp. 1342–1345. [CrossRef]

32. Nirschl, J.J.; Janowczyk, A.; Peyster, E.G.; Frank, R.; Margulies, K.B.; Feldman, M.D.; Madabhushi, A. A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H&E tissue. *PLoS ONE* **2018**, *13*, e0192726.

33. Rao, S.; Li, Y.; Ramakrishnan, R.; Hassaine, A.; Canoy, D.; Cleland, J.; Lukasiewicz, T.; Salimi-Khorshidi, G.; Rahimi, K. An explainable Transformer-based deep learning model for the prediction of incident heart failure. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 3362–3372. [CrossRef]

34. Gjoreski, M.; Gradišek, A.; Budna, B.; Gams, M.; Poglajen, G. Machine learning and end-to-end deep learning for the detection of chronic heart failure from heart sounds. *IEEE Access* **2020**, *8*, 20313–20324. [CrossRef]

35. Pana, M.A.; Busnatu, S.S.; Serbanoiu, L.I.; Vasilescu, E.; Popescu, N.; Andrei, C.; Sinescu, C.J. Reducing the Heart Failure Burden in Romania by Predicting Congestive Heart Failure Using Artificial Intelligence: Proof of Concept. *Appl. Sci.* **2021**, *11*, 11728. [CrossRef]

36. D'Addio, G.; Donisi, L.; Cesarelli, G.; Amitrano, F.; Coccia, A.; La Rovere, M.T.; Ricciardi, C. Extracting Features from Poincaré Plots to Distinguish Congestive Heart Failure Patients According to NYHA Classes. *Bioengineering* **2021**, *8*, 138. [CrossRef]

37. Kwon, J.m.; Kim, K.H.; Jeon, K.H.; Kim, H.M.; Kim, M.J.; Lim, S.M.; Song, P.S.; Park, J.; Choi, R.K.; Oh, B.H. Development and validation of deep-learning algorithm for electrocardiography-based heart failure identification. *Korean Circ. J.* **2019**, *49*, 629–639. [CrossRef]

38. Porumb, M.; Iadanza, E.; Massaro, S.; Pecchia, L. A convolutional neural network approach to detect congestive heart failure. *Biomed. Signal Process. Control.* **2020**, *55*, 101597. [CrossRef]

39. Li, D.; Tao, Y.; Zhao, J.; Wu, H. Classification of congestive heart failure from ECG segments with a multi-scale residual network. *Symmetry* **2020**, *12*, 2019. [CrossRef]

40. Chicco, D.; Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 16. [CrossRef] [PubMed]

41. UCI. Heart Disease Data Set. Available online: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data (accessed on 1 November 2022).

42. Ishaq, A.; Sadiq, S.; Umer, M.; Ullah, S.; Mirjalili, S.; Rupapara, V.; Nappi, M. Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. *IEEE Access* **2021**, *9*, 39707–39716. [CrossRef]

43. Ghosh, P.; Azam, S.; Jonkman, M.; Karim, A.; Shamrat, F.J.M.; Ignatious, E.; Shultana, S.; Beeravolu, A.R.; De Boer, F. Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access* **2021**, *9*, 19304–19326. [CrossRef]

44. Guo, A.; Pasque, M.; Loh, F.; Mann, D.L.; Payne, P.R. Heart failure diagnosis, readmission, and mortality prediction using machine learning and artificial intelligence models. *Curr. Epidemiol. Rep.* **2020**, *7*, 212–219. [CrossRef]

45. Alotaibi, F.S. Implementation of machine learning model to predict heart failure disease. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 261–268. [CrossRef]

46. Aljanabi, M.; Qutqut, M.H.; Hijjawi, M. Machine learning classification techniques for heart disease prediction: A review. *Int. J. Eng. Technol.* **2018**, *7*, 5373–5379.

47. Kannan, R.; Vasanthi, V. Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease. In *Soft Computing and Medical Bioinformatics*; Springer: Singapore 2019; pp. 63–72.

48. Sudoh, T.; Kangawa, K.; Minamino, N.; Matsuo, H. A new natriuretic peptide in porcine brain. *Nature* **1988**, *332*, 78–81. [CrossRef]

49. Lee, Y.K.; Choi, D.O.; Kim, G.Y. Development of a Rapid Diagnostic Kit for Congestive Heart Failure Using Recombinant NT-proBNP Antigen. *Medicina* **2021**, *57*, 751. [CrossRef]

50. Kasahara, S.; Sakata, Y.; Nochioka, K.; Miura, M.; Abe, R.; Sato, M.; Aoyanagi, H.; Fujihashi, T.; Yamanaka, S.; Shiroto, T.; et al. Conversion formula from B-type natriuretic peptide to N-terminal proBNP values in patients with cardiovascular diseases. *Int. J. Cardiol.* **2019**, *280*, 184–189. [CrossRef] [PubMed]

51. Nabeshima, Y.; Sakanishi, Y.; Otani, K.; Higa, Y.; Honda, M.; Otsuji, Y.; Takeuchi, M. Estimation of B-type natriuretic peptide values from n-terminal proBNP levels. *J. UOEH* **2020**, *42*, 1–12. [CrossRef] [PubMed]

52. Cameron, S.J.; Green, G.B.; White, C.N.; Laterza, O.F.; Clarke, W.; Kim, H.; Sokoll, L.J. Assessment of BNP and NT-proBNP in emergency department patients presenting with suspected acute coronary syndromes. *Clin. Biochem.* **2006**, *39*, 11–18. [CrossRef] [PubMed]

53. Zhao, J.; Feng, Q.P.; Wu, P.; Lupu, R.A.; Wilke, R.A.; Wells, Q.S.; Denny, J.C.; Wei, W.Q. Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction. *Sci. Rep.* **2019**, *9*, 717. [CrossRef]

54. Rongali, S.; Rose, A.J.; McManus, D.D.; Bajracharya, A.S.; Kapoor, A.; Granillo, E.; Yu, H. Learning latent space representations to predict patient outcomes: Model development and validation. *J. Med Internet Res.* **2020**, *22*, e16374. [CrossRef]

55. Golas, S.B.; Shibahara, T.; Agboola, S.; Otaki, H.; Sato, J.; Nakae, T.; Hisamitsu, T.; Kojima, G.; Felsted, J.; Kakarmath, S.; et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: A retrospective analysis of electronic medical records data. *BMC Med. Inform. Decis. Mak.* **2018**, *18*, 44. [CrossRef]

56. Choi, E.; Schuetz, A.; Stewart, W.F.; Sun, J. Using recurrent neural network models for early detection of heart failure onset. *J. Am. Med Inform. Assoc.* **2017**, *24*, 361–370. [CrossRef]

57. Gron, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1st ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2017.

58. Chollet, F. *Deep learning with Python*; Simon and Schuster: New York, NY, USA, 2021.

59. Dahouda, M.K.; Joe, I. A Deep-Learned Embedding Technique for Categorical Features Encoding. *IEEE Access* **2021**, *9*, 114381–114391. [CrossRef]

60. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.

61. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

62. Guidi, G.; Pollonini, L.; Dacso, C.C.; Iadanza, E. A multi-layer monitoring system for clinical management of Congestive Heart Failure. *BMC Med. Inform. Decis. Mak.* **2015**, *15*, S5. [CrossRef] [PubMed]

63. Pham, T.; Tran, T.; Phung, D.; Venkatesh, S. Predicting healthcare trajectories from medical records: A deep learning approach. *J. Biomed. Inform.* **2017**, *69*, 218–229. [CrossRef] [PubMed]

64. Lu, X.H.; Liu, A.; Fuh, S.C.; Lian, Y.; Guo, L.; Yang, Y.; Marelli, A.; Li, Y. Recurrent disease progression networks for modelling risk trajectory of heart failure. *PLoS ONE* **2021**, *16*, e0245177. [CrossRef] [PubMed]

65. Guo, A.; Beheshti, R.; Khan, Y.M.; Langabeer, J.R.; Foraker, R.E. Predicting cardiovascular health trajectories in time-series electronic health records with LSTM models. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 5. [CrossRef]