
Deep Transfer Learning Applied to Time-series Classification for Predicting Heart Failure Worsening Using Electrocardiography

A dissertation

Submitted to the faculty of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirement for the Degree of

Master of Science in

Biomedical Engineering

June 7th, 2020

by

Xiang Pan

Approved:

Kwonmoo Lee, Ph. D.

Assistant Professor, Advisor

Department of Biomedical Engineering

Worcester Polytechnic Institute

Songbai Ji, Ph. D.

Associate Professor, Committee Chair

Department of Biomedical Engineering

Worcester Polytechnic Institute

Yitzhak Mendelson, Ph. D.

Professor

Department of Biomedical Engineering

Abstract

Computational ECG (electrocardiogram) analysis enables accurate and faster diagnosis and early prediction of heart failure related symptoms (heart failure worsening). Machine learning, particularly deep learning, has been applied for ECG data successfully. The previous applications, however, either mainly focused on classifying occurrent, known patterns of on-going heart failure or heart failure related diseases such arrhythmia, which have undesirable predictability beforehand, or emphasizing on data from pre-processed public database data. In this dissertation, we developed an approach, however, does not fully capitalize on the potential of deep learning, which directly learns important features from raw input data without relying on a priori knowledge. Here, we present a deep transfer learning pipeline which combines an image-based pretrained deep neural network model with manifold learning to predict the precursors of heart failure (heart failure-worsening and recurrent heart failure related re-hospitalization) using raw ECG time series from wearable devices.

In this dissertation, we used the unprocessed real-life ECG data from the SENTINEL-HF study by Dovancescu, et al. to predict the precursors of heart failure worsening. To extract rich features from ECG time series, we took a deep transfer learning approach where 1D time-series of five heartbeats were transformed to 2D images by Gramian Angular Summation Field (GASF) and then the pretrained models, VGG19 were used for feature extraction. Then, we applied UMAP (Uniform Manifold Approximation and Projection) to capture the manifold of the standardized feature space and reduce the dimension, followed by SVM (Support Vector Machine) training. Using our pipeline, we demonstrated that our classifier was able to predict heart failure worsening with 92.1% accuracy, 92.9% precision, 92.6% recall and F1 score of 0.93 bypassing the detection of known abnormal ECG patterns.

In conclusion, we demonstrate the feasibility of early alerts of heart failure by predicting the precursor of heart failure worsening based on raw ECG signals. We expected that our approached provided an innovative method to assess the recovery and successfulness for the treatment patient received during the first hospitalization, to predict whether recurrent heart failure is likely to occur, and to evaluate whether the patient should be discharged.

Acknowledgements

I still remember six years ago, in 2014, when I first walked into this great campus with empty head. It still feels like yesterday. However, I am not the same person six years ago. With all the brilliant professors and extraordinary classmates, this campus had taught me so much that I can be benefited for a lifetime. Back six years ago, I would never imagine that I can have that much knowledge to finish this dissertation and work in a thesis that might be able to change the world. There are several people that I would like to show my greatest gratitude to, without them I would still be the same child with empty head and do not know where my future is.

First of all, I would like to show my highest gratitude to my advisor, Professor Kwonmoo Lee, who has been a great mentor and pioneer to me. Over the past two years, you have been supporting my back and giving me tremendous opportunities. From the early learning experience of machine learning to all the discussion and deliberation, you are always ready to give your full support to help me to understand. During this thesis, you taught me the difference in all sorts of machine learning techniques and showed me how to establish a robust structure with the best performance. I still remember when my initial attempt over the project went into a dead end, I felt deeply confused and depressed from the obstacle in front of me, it was you who straightened my thinking and guided me back to the right way with his experience. You also taught me how to understand an unanticipated result with reasonable explanation, how to illustrate with more concise language in paper writing, and not to give up when struggling. With your help, I learned how to make assumption fearlessly but seek validation carefully. I am deeply impressed and inspired by your enthusiasm and dedication in learning and developing machine learning algorithm.

I also appreciate the help from my committee members with the constructive comments and advices. I would like to thank Professor Songbai Ji, whose comment always go right to the heart of the matter and point out where the limitation of the approach is. Without his help, I would not be able to implement the structure and workflow to achieve better performance. I would like to thank Professor Yitzhak Mendelson, who not only taught me how to interpret signals from biomedical instruments, but also introduced and provided this dataset for me to study and use. When I have questions about the labeling of the dataset, he spared no effort to helped me find the interpretation. I would also like to thank Professor Ki Chon, who is not my committee member, but brought tremendous, valuable

comments reviewing the manuscript.

Beyond this, I would like to thank the Laboratory of Quantitative Cell Imaging and lab members. They have always been enthusiastic when I search for comment, and I learned a lot from them. I want to thank Dr. Heejune Choi and Dr. Chuangqi Wang, who not only inspired me with their work, but also taught me how to organize my language in presentation. I would also like to thank Yudong Yu, Junbong Jang, Pengyi Ye, Mengzhi Cao, Xinye Fan, and Dr. Tzu-Hsi Song for their valuable comments.

At last but certainly not least, I would like my parents, who has been supporting me and be my biggest fans for 24 years. You might not know but I have been missing you all the time when I am away from you. I love you and really wish you can be healthy and happy all the time. I would also like to thank my wonderful friends who has been with me the entire time when I am at WPI. You were always here with me with all your encouragement.

Content

Figures	ii
Tables.....	ii
Chapter 1: Introduction	1
1.1 The develop of machine learning.....	1
1.2 Biological background of acute decompensated heart failure (ADHF), diagnosis, treatment, and consideration after discharge from hospitalization	2
1.3 Problem statement and strategy	3
1.4 Application of deep learning and transfer learning on heart failure diagnosis and previous studies.....	4
1.5 Application of SENTINEL-HF dataset in this study	6
1.6 Manifold learning and dimensionality reduction techniques	9
Chapter 2: Materials and Methods.....	11
2.1 Raw data preparation, five heart-beat structure remodeling and sample selection.....	11
2.2 Early data identification and class labeling	12
2.3 Transformation from 1D time series into 2D images	14
2.4 Feature extraction with VGG19 pre-trained neural network	14
2.5 Multilayer perceptron neural network classification	15
2.6 Standardization for sample individuality removal	16
2.7 Training classifier with support vector machine (SVM).....	17
2.8 Hyperparameter testing	17
2.9 Comparative methods as alternatives.....	19
Chapter 3: Results	21
3.1 PCA dimensional reduction and 2D visualization with t-SNE on the reduced features	21
3.2 Initial Classification based on MLP neural network classifier	22
3.3 Balancing global and local data structures by UMAP	23
3.4 Implemented new pipeline with standardization, manifold learning of UMAP and SVM classifier	23
3.5 Roles of each computational step in our pipeline.....	24
Chapter 4: Conclusion and outlook.....	29
References.....	34

Figures

Figure 1: Raw data of five minutes ECG samples from SENTINEL-HF study	7
Figure 2: Example of excluded data that has not been used in this study	8
Figure 3: Comparison of raw ECG data and transformed image exhibit sample individuality	10
Figure 4: Remodeling of five minutes ECG time series into five heartbeat ECG time series.....	13
Figure 5: Distribution of data and source subjects among Class HF and N	13
Figure 6: Transformation of ECG time series into 2D images by GASF	14
Figure 7: Schematics of VGG19 model we used for feature extraction	16
Figure 8: Visualization of initial attempt of the pipeline with default parameters in UMAP	18
Figure 9: Example of initial step of parameter searching	19
Figure 10: Schematics of image transformation process of a five heartbeat ECG time series into 2D	20
Figure 11: Cumulative explained variance ratio (EVR) and explained variance ratio plot	21
Figure 12: t-SNE visualization on 2D scale with class labels and patient labels	22
Figure 13: Balancing global and local data structure by UMAP	25
Figure 14: Classification performance on feature extracted from early prediction of heart failure worsening ECG data.....	27
Figure 15: 2D t-sne visualization on No_UMAP	28
Figure 16: 2D UMAP visualization on No_STD	28
Figure 17: Classification performance comparison among different image transformation methods (GASF, MTF and RP).....	29
Figure 18: Superior workflow of the application of this methodology on other dataset and studies..	33

Tables

Table 1: Optimum classification performance main approach and alternative pipelines achieved and corresponding parameters	26
---	----

Chapter 1: Introduction

1.1 The develop of machine learning

The history of machine learning and artificial intelligence can be tracked back to the 1950s, when Arthur Samuel was working at International Business Machines Corporation (IBM), he brought up, and later examined the idea that computer algorithms can be programed to learn from itself and improve with better performance with continuously training¹. Unfortunately, until the day he passed away, at July 29th 1990, the development of theories and algorithms in machine learning had been promoting in a sluggish pace. In the contemporary age, particularly in the past two decades, the merits of machine learning have been paid much more attention under the spotlight. Applications and improvements sprang up like mushrooms in various fields and all sorts of studies. Instead of existed only in textbooks and theories, the concept of machine learning and artificial intelligence have been introduced to public and started influencing people's daily life. Recent progress in machine learning, particularly deep learning has shown that, in fields that involves analysis of complex, large and/or high dimensional dataset, such as image classification, audio recognition, and games, computer (artificial intelligence) has the ability to outperform human, and the application of machine learning and artificial intelligence have significantly increased people's quality of life²⁻⁸. It not only brought benefits to people's daily life, but also changed the opinion of entertainment. In 2016, DeepMind Technologies introduced AlphaGo, a deep learning based Go game artificial intelligence, which later successively defeated Korean chess player Lee Se-dol and Chinese chess player Ke Jie (two of the top human level chessmen)⁹⁻¹¹. In 2018, OpenAI Five was developed and rewrote the history in online video game DOTA 2, defeating 99.4% of human players, including world championship team, in over 7000 games¹².

In another way, in the field of medical research and healthcare, machine learning has shown significant potentials in the areas such as medical diagnosis, diagnostic reasoning and signal processing¹³⁻¹⁷. Compare with human expertise, computational analysis has extraordinary performance, and meanwhile has huge advantages in portability, efficiency dealing with large amount of data, and convenience. Deep neural networks (DNN), particularly convolutional neural networks (CNN)^{18,19} have shown their higher feasibility in medical applications in which diagnosis consists of observation of images, such as breast cancer classification and lung area detection with unique

advantages^{20,21}. With the technique of machine learning, diagnosis of diseases becomes possible for people in areas with low allocation of medical resources. The development of machine learning is increasing people's quality of life. However, in some fields, due to the fact that machine learning lacks human expertise' breadth in experience, adjustability and knowledge, it still remains huge area of improvement. In novel studies such with data that hasn't been trained yet previously, some abnormalities can be difficult to classify for machine learning

1.2 Biological background of acute decompensated heart failure (ADHF), diagnosis, treatment, and consideration after discharge from hospitalization

Acute decompensated heart failure (ADHF) is one of the major types of heart failure with cardiovascular worsening syndrome, triggered by insufficient pump of blood from heart to organs, and causes serious symptoms such as difficulty in breathing and organ failure²². It is known as a complication of other cardiovascular diseases, including myocardial infarction, abnormal heart rhythm and infection²³. In the contemporary age, ADHF has been attracting more and more public attention, for the growing trend of having ADHF and annual increase in correlated mortality. According to national hospital discharge survey, in United States, the hospitalization of ADHF has been tripled from 1,274,000 in 1979 to 3,860,000 in 2004, while approximately 20 million people is affected by ADHF worldwide by the year of 2009²⁴. Although the increase in the number of patients receiving treatment also represents the development of medical technique in diagnosis, disease screening and prevention, it is widely believed that the unhealthy lifestyle of people is one of the main causes of ACHF. While busy working/studying and neglecting the importance of exercising, modern people have a larger chance to be negatively influenced by long-term stress (reduction of sleep and sleep quality)^{25,26}, increasing in body weight (obesity)²⁷ and excessive sodium intake (high blood pressure)²⁸.

The outcome of attentionless to ADHF is life-threatening, but with proper, on-time treatment the risk can still be minimized. Cardiologists often carry out first-line therapies with decongestion approaches, such as using intravenous loop diuretics (for example, furosemide, bumetadine, and torsemide) to increase urine output and ease the syndrome²⁹. Although the effectiveness of treatment often suggesting promising result during evaluation for discharge, recent study suggested that recurrent congestion, which could lead to a second HF, is still very high (51% and 27% for moderate

and severe congestion in the first month after discharge)³⁰, the mortality rate after discharge is approximately 20%^{31,32}. But even after receiving proper treatment, depending on the successfulness of treatment patients received, such as decongestion, during hospitalization, patients discharged from hospital can still remain high possibility to experience recurrent congestion, which could further lead to a second ADHF³⁰. Therefore, it is important to accurately evaluate and predict the likelihood of recurrent heart failure worsening during and after discharge from hospital.

1.3 Problem statement and strategy

This study mainly focused on developing a methodology to predict whether it is feasible to accurately predict recurrent heart failure worsening. The main focus in this thesis is to develop a methodology that can most accurately predict beforehand, so that patient can be warned and have enough time to receive additional treatment when a second heart failure recur. In addition to the performance and accuracy of the approach, there are several other aspects that have been taken into consideration during this thesis:

- 1) Whether it is possible to develop a system, so that patients discharged from hospital can take the measurement at home on their own, instead of visiting hospital for daily/routine evaluation?**
- 2) Whether current public dataset can fulfill the data demand for this project?**
- 3) Whether predict on heart failure worsening or classify on on-going heart failure patterns**

For people in rich, urban area, medical resources are adequately allocated. But for people in desolated regions or developing counties, regular examination and evaluation is relatively more difficult to be achieve. Therefore, it is important the examination/evaluation can be carried out by patients themselves when they are at home. It further derives two requirements. First of all, the procedure for data collection should be simple and user-friendly, so that people who carry out the procedure does not require a long learning process and can accurately acquire data without systematic errors. Second, the procedure that leads to the final result should be easily accessible without complex tests. Traditional tests for heart failure diagnosis, such as blood test, requires high medical and laboratory background and requisition for equipment and materials. Based on these two

requirements, I expected that computer aid calculation, more specifically approach involves machine learning could be a good media to carry out the evaluation. For one, computer aided calculation can be easily achieved remotely with the developed technology in the past decade. Meanwhile the popularization of internet makes the update in training set feasible. For another, with proper administration, the feedback from individual patient can enlarge the dataset and makes future evaluation more accurate. On the other hand, electrocardiogram (ECG) can be easily acquired with electrodes and used as the time series input for machine learning models, which satisfies both requirements. Therefore, ECG has been used as the main target for this thesis.

This thesis mainly focused on testing the feasibility of using machine learning techniques to distinguish the precursor of recurrent heart failure worsening. However, almost all of current public datasets share some common drawbacks. First of all, due to the fact that many studies value more in the determination of heart failure when it occurs, the majority of public dataset contains abnormal heartbeat of ECG during heart failure. And because observation of ECG is also taken as one of diagnosis tool, these public datasets were evaluated and labeled by cardiologists, and often exhibit abnormality/patterns that can be visually distinguished. The lack of data beforehand makes it difficult to restore the scenario of before HF takes place. Furthermore, majority of public dataset included data when patients were diagnosing whether they had HF. However, the data that after patients discharged from hospital were not available. Therefore, completeness and relevance have been considered as two main reasons we focused the real-life dataset of SENTINEL-HF study instead of other public databases as primary target of this thesis. On the other aspect in the clinical perspective, prediction of heart failure worsening, compared with classification on on-going HF pattern, has the advantage of better prevention in mortality of heart failure and correlated diseases

1.4 Application of deep learning and transfer learning on heart failure diagnosis and previous studies

In the past decade, the study related to application of deep learning on heart failure diagnosis has been attracting public's attention. In fact, the process of traditional diagnosis of heart failure can be time-consuming. It normally started with patients feel discomfort, they visit hospital and cardiologist. Depending on different methods been taken (such as echocardiogram and blood tests) and urgency of the situation, the time span from the beginning, when patients start feeling discomfort, to diagnosis

result varies from hours to days. Electrocardiogram (ECG) is one of the parameters frequently being used in diagnosis of heart failure and correlated abnormalities. Cardiologists use the observation of unique ECG features, such as abnormalities presented by atrial fibrillation, myocardial infarction or arrhythmia, as one of the factors to carry out clinical diagnosis of myocardial ischemia³³⁻³⁵. Compared to these ordinary methods to diagnose heart diseases, computational ECG analysis provides great advantages including accurate and faster diagnosis and early prediction of heart failure^{36,37}. Together with remote monitoring using wearable devices, it also provides new opportunities for telemedicine. However, existing algorithms for diagnosis of heart diseases leave ample room for improvement in prediction accuracy and due to limited testing with real-world dataset³⁸⁻⁴².

Since the beginning of 2010s, more and more researches have been focusing on cardiovascular-related classification with ECG as input in the study of machine learning. It has been proved that when dealing with certain types of cardiovascular disease, machine learning can give accurate prediction, mostly based with known pattern on ECG^{43,44}. In order to create model with better in performance, researchers have applied numerous approaches. Data collection with different time span were applied for evaluation under different circumstances. Different data treatment technique (for example, dimensional reduction and denoising) has been developed. Moreover, deep learning models with different structures and data mining techniques were introduced and compared to enhance the performance of classification. To show the advantage of the model in performance over previous models, some of the existed studies used previously published datasets, which are known as public database, as the source of data for deep learning pipelines. By using data from known public database, such as MIT-BIH Arrhythmia database⁴⁵ and BIDMC congestive Heart Failure database⁴⁶, researchers are able to develop algorithms for different purposes, for example image recognition and pattern detection^{45,47,48}. It also provided them a platform to compare and show the robustness and performance of their models with other existed models^{45,49-51}, so that the value of the new techniques can be emphasized.

Although DNN also has been applied for ECG data successfully, the previous applications were mainly focused on classifying known patterns of arrhythmia⁵². While the accurate and automatic detection of abnormal ECG patterns will greatly aid healthcare decisions, this approach does not fully capitalize on the potential of deep learning, which directly learns important features from raw input data without relying on a priori knowledge. Recent studies suggest that the patients' risk of

cardiovascular death⁵³ or congestive heart failure⁵⁴ and cardiac contractile dysfunction⁵⁵ can be predicted by analyzing raw ECG signals. However, the early prediction of the precursors of heart failure using raw ECG signals has not been demonstrated.

Convolutional neural network (CNN) is one type of deep neural network (DNN) that contains convolution calculation⁵⁶. Due to the fact that it has representation learning ability to carries out shift-invariant classification to input information, CNN is also referred as Shift-Invariant Artificial Neural Networks (SIANN)⁵⁷. Compare with other DNN structures with the same number of hidden units, CNN has fewer hyperparameters, thus makes it easier to train. Since in 1995, LeCun, et al. applied CNN on handwritten digit recognition⁵⁸, more and more studies proved that convolutional neural network outperforms simple neural networks in many fields, including image recognition^{59,60}, natural language processing⁶¹ and so on.

Transfer learning has been widely adopted in deep learning practices to build robust classifiers mainly with image datasets⁶². Particularly, the CNN models pretrained with large numbers of natural images available in the ImageNet⁶³ were used as general-purpose image feature extractors to build the classifiers in different domains⁶⁴⁻⁶⁸. Furthermore, in the studies with limited dataset, applying transfer learning has been proved an effective way to reduce overfitting⁶⁹. The effective pretrained model for time series data, however, is still limited, since no large-scale time series datasets is available, making it difficult to extract rich features from time series data. VGGNet is a pre-trained model that takes merits of transfer learning, recently proposed by Simonyan, et al⁷⁰. Different from traditional CNN models, VGGNet used 3×3 convolution kernels and 2×2 max pooling. For one, this distinctive structure provides better learning ability on features. For another, it further reduces the number of hyperparameters. In this study we used VGG19, a VGGNet that contains 19 layers (16 convolutional layers and 3 fully-connected layers).

1.5 Application of SENTINEL-HF dataset in this study

In this study, we were looking for a dataset that we can used to study about the risk of recurrent heart failure and related heart failure worsening conditions. The dataset should cover the time span from the time that patients discharged from the hospital, to the time that worsening condition shows up (if existed). However, due to the lack of study in the subject, exist public databases are not the best option to be used in this study. In this circumstance, we used an unprocessed real-life dataset

from SENTINEL-HF study⁷¹. In Dovancescu, S. et al. original study, they targeted at transthoracic bioimpedance as main subject, and they found that it can be used as an indicator of heart failure decompensation. During the study, they designed a vest that contains four electrodes that can be used to collect data when people wear it. In addition to transthoracic bioimpedance, the vest also produces one channel ECG recordings as output. After discharged from hospital from ACHF, patients took routine/daily measurement of transthoracic bioimpedance by wearing the vest. Every time each person used the equipment, an approximately five minutes single channel ECG were collected. The time span of the study started from the day patients discharged from the hospital after receiving treatment of ACHF, to the day heart failure worsening recurred (if possible), or for more than 45 days if the patient stayed healthy and recovered into better health condition from ACHF. In view of these aspects, SENTINEL-HF dataset perfectly match the requirement for this thesis. These single channel ECG time series were used as our raw data in this study (Figure 1).

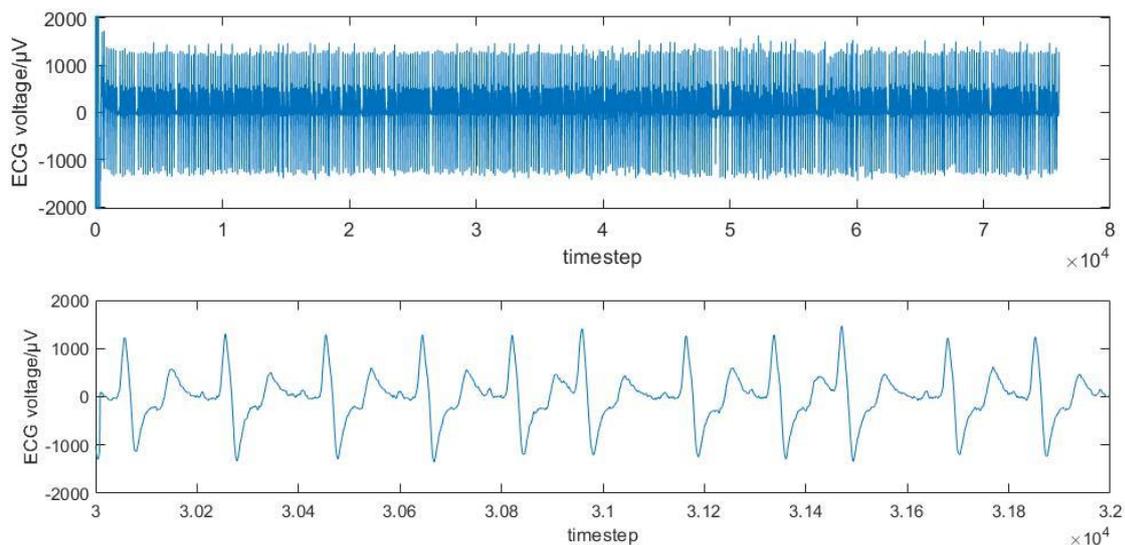


Figure 1: Raw data of five minutes ECG samples from SENTINEL-HF study. Upper: ECG voltage over the whole timespan during a piece of early (right after discharged) recording from patients that did not experience heart failure worsening during the study. Lower: magnified plot of raw ECG data from the same recording.

Not all of the ECG data from SENTINEL-HF study was being used in this study. We carried out a manual data selection and remodeling step. First of all, data from some of the patients were labeled “abandoned” in previous SENTINEL-HF study. The reason for these samples being labeled can be complicated. For example, it can be either the result of subject quit the program, left incompetence in his/her sub-dataset, or the health condition of the patient was not well studied. For either reason, we

did not use data from the patients whose data had been labeled as “abandoned”. Next, due to the error from equipment, the ECG recordings from some patients can be difficult to be interpreted. As shown in Figure 2, large noise covered the whole timespan and the pattern of ECG could not be identified. The cause of this phenomenon can be either component damage in the equipment or patients did not use the equipment in the correct way. Therefore, after acquiring the raw dataset from SENTINEL-HF study, all the recordings had been examined and data that was unusable to this thesis was excluded.

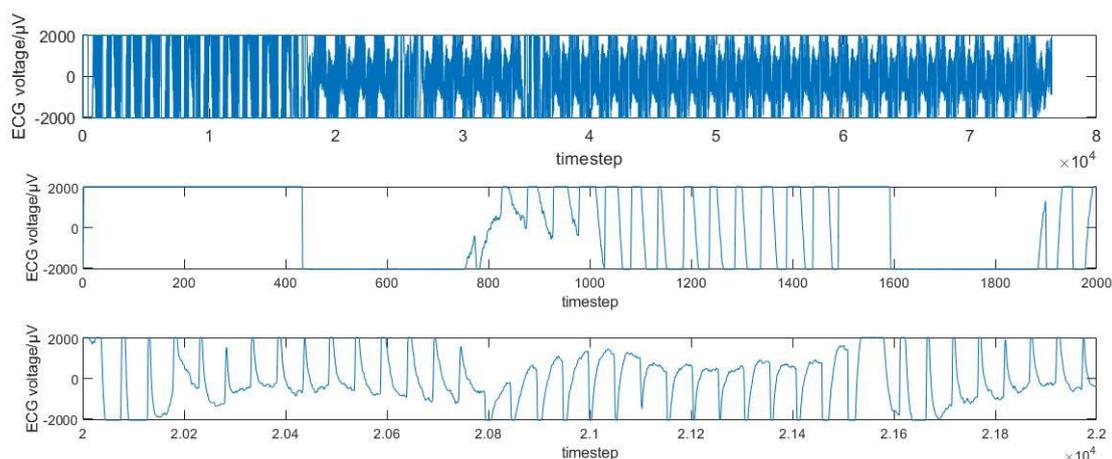


Figure 2: Example of excluded data that has not been used in this study (the recording was inundated with huge noise and un-interpretable pattern).

In this study, we did not directly use five-minute ECG recordings as our input model. First of all, as the sensitivity of the electrodes also provided unwanted noise when patients move the part of body and interfere the measurements. This kind of noise often appears at the beginning and at the end of the recording (as shown in Figure 1 as well), however sometimes it also exists at the middle of a recording. For another, after removing unusable data, the total sample size of five-minutes ECG time series is limited. Moreover, five minutes ECG contains unwanted variables that might affect the classification result. For example, the heart rate and exact time length for each recording differs a lot. In order to take into consistency, we used a customized strategy to remodel samples into five heartbeat ECG time series. The five cycles form not only enlarged the sample size, but also created the advantage that the transition pattern of ECG (both the pattern from QRS interval to the next QRS interval, and from T wave to the next T wave) between heartbeats can be taken into consideration. Many public database data, for example ECG 200⁷², contains only single heartbeat ECG time series.

Among the dataset that covers the entire time span, the earliest data that contains patients' ECG within the time span of short time after they discharged from hospital were focused as the main target for this thesis. These earliest data were then labeled as (1) early heart failure worsening indicator (class HF) or (2) normal (class N) in accordance with the health status of the patient during the main SENTINEL-HF study and corresponding followed up period (see Materials and Method for details).

The application of unprocessed real-life dataset can sometimes be limited by imperfection of data collected. As shown in Figure 3, raw ECG data in SENTINEL-HF study also exhibited difference in the pattern between different equipment and subjects. Morphologically speaking, same pattern existed in all recording from the same patient, and can hardly be eliminated by pre-processing of raw data (such as normalization). Due to the fact that the data collection process requires patient's manual operation (wear the device), the ECG data, recorded by the electrodes, were often not collected with unified standard. Other possible cause can be the difference in individual equipment. But on the other hand, imperfection is one of the disadvantages of unprocessed real-life dataset which we had forecasted. The data in majority public dataset were manually selected, many of which had been pre-processed with techniques such as interpolation and resizing⁷². Unlike other public datasets, SENTINEL-HF dataset has unpredictability in different situations. Therefore, pre-processing and post-processing for the raw data and features are extremely important before applying classifier for prediction. Meanwhile, the workflow must be robust enough to overcome the imperfection of the samples.

1.6 Manifold learning and dimensionality reduction techniques

In the idea of manifold learning, high-dimensional data can often be embedded into low-dimensional form for ease of observation⁷³. Due to the limitation within the high dimensional structure, the redundant of dimensionality can exist. Supervised and unsupervised linear dimensionality reduction frameworks, such as PCA, transform components from high-dimensional form into corresponding low-dimensional vector by linear function/relationship⁷⁴. However manifold learning uses another strategy, non-linear dimensionality reduction technique, in which the low-dimensional output components are given by non-linear function of the components of high-dimensional input⁷⁵. Compare with linear dimensionality reduction, manifold learning is more sensitive to non-linear structure in data, thus creates better preservation in local structure of the dataset, which reflect the

interpretation of distance between points within the same clusters⁷³. Like linear dimensional reduction frameworks, manifold learning can be either supervised or unsupervised. In this thesis we used unsupervised approach, in which the model learned from the high-dimensional structure of the data by itself without providing pre-determined label.

A Five heartbeats ECG time series examples **B** Corresponding GASF 2D image examples

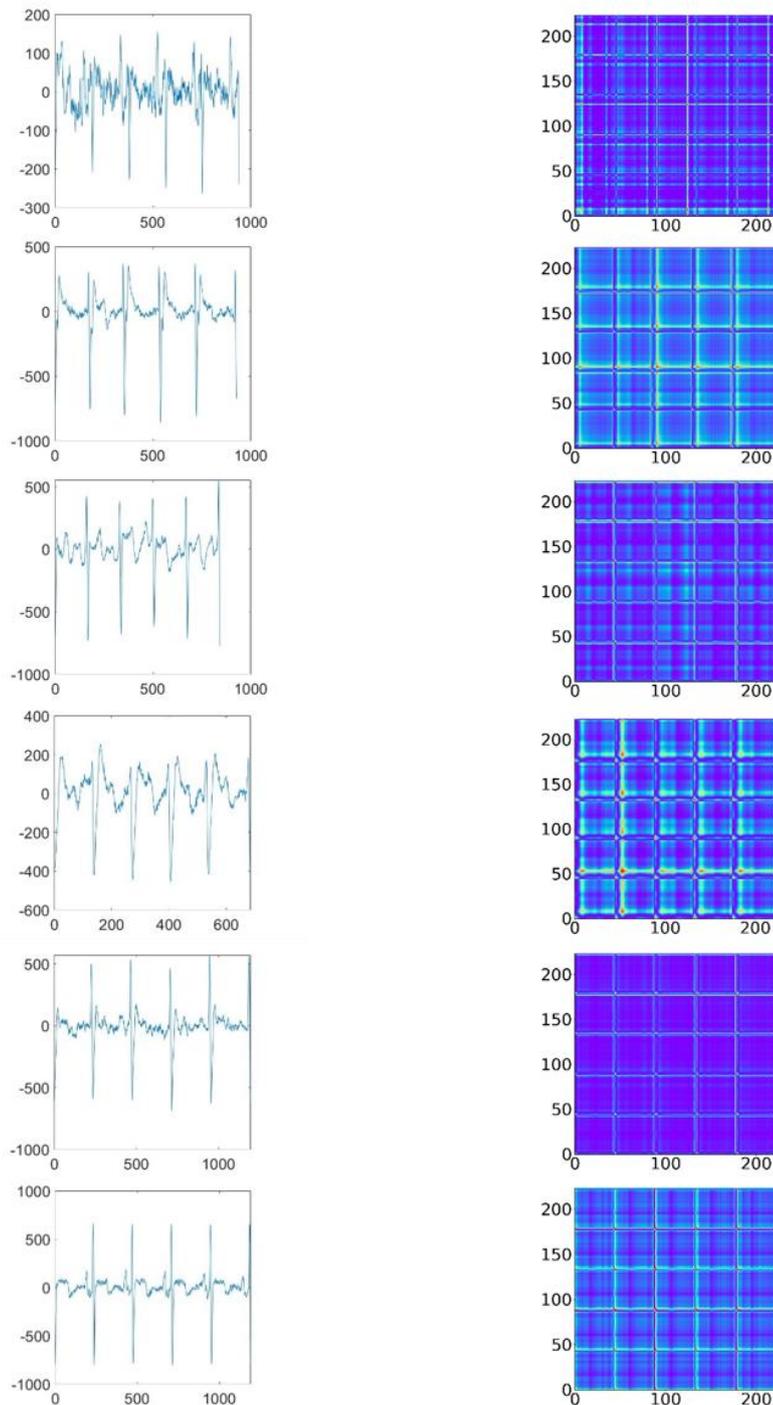


Figure 3: Comparison of raw ECG data and transformed image exhibit sample individuality that is observable, applied rainbow colormap for ease of visualization

Traditional manifold learning frameworks include Isomap, locally linear Embedding (LLE), modified locally linear embedding (MLLE), t-SNE (which we also previously used to visualize the VGG19-PCA features) and so on, has been widely applied in different applications that requires dimensional reduction to enhance performance and reduce computational time⁷⁶⁻⁸¹. In this thesis, we selected a different framework, Uniform Manifold Approximation and Projection (UMAP). Compared with t-SNE, one of the other manifold learning frameworks that has been most widely used in all sorts of applications, UMAP is more time-efficient, and preserves not only data's local structure, but also the global structure of the dataset⁸². While local structure shows the difference between data within the same type with different distance between points in the same cluster, global structure reflects the difference between different clusters with distance between different clusters⁸³. By controlling one important parameter $n_neighbors$ in UMAP, different of neighboring points can be connected when constructing the low-dimensional representation of the high-dimensional structure. In order to test the effect of global and local structure preservation on classification performance, we applied grid searching technique over $n_neighbors$ and min_dist . In that way, the balance between global and local structure preservation can be achieved.

Chapter 2: Materials and Methods

2.1 Raw data preparation, five heart-beat structure remodeling and sample selection

SENTINEL-HF is a prospective study of patients discharged after hospitalization for acute decompensated heart failure (ADHF) associated with sudden or gradual worsening of heart failure symptoms²³. The original complete dataset from SENTINEL-HF study contains ECG recordings from 82 different patients. However, not all of the data was being used in this thesis. We carried out data selection step twice. The primary data selection excluded raw data that was labeled with "abandoned" or visually inspected as uninterpretable. After the primary data selection, we applied a customized five heartbeat ECG extraction method, as shown in Figure 4. A peak detection algorithm was used to record S wave locations in MATLAB (MathWorks), by extracting out the location of the minimum of the valley that is less than 95% of the data. We used the MATLAB built-in function *findpeaks* with 5% quantile as '*MinPeakHeight*' and 100 as '*MinPeakDistance*'. At each location of S wave in ECG signal, sub-segments with five consecutive heartbeats were extracted. The followed-up sub-segment started

consecutively from one heartbeat behind the start of the previous sub-segment. During the training set preparation, we visually inspected the ECG data and removed data with artifacts (including large noises and missing/extra parts). At the end, we selected 4415 five heartbeats ECG time series among the whole-time span in SENTINEL-HF study from 32 patients.

2.2 Early data identification and class labeling

We selected the earliest ECG data (from the first 1-5 recorded days depending on data availability) from the patients who experienced later heart failure or heart failure worsening indicators during Dovancescu, et al. study (include follow-up period) as the precursor of heart failure worsening⁷¹. For the patients who were not rehospitalized or experience HF-worsening within the span of the study, we selected data from earliest five recordings as the normal group. In this way, ECG segments were categorized into two classes:

- Class HF (Rehospitalized, HF-worsening or HF): The early data from patients who later experienced HF-worsening and/or cardiac ischemic-related heart failure during the study and follow-up period.
- Class N: The data from the earliest five days of patients who were not rehospitalized or experience HF-worsening during the study and follow-up period.

We verified 275 five heartbeats ECG time series. The samples were acquired from the previously selected dataset for the entire timespan. It contains data from 29 patients. Early data from heart failure worsening (HF) contains 150 ECG time series (from 18 different subjects), and the normal data (N) contains 125 ECG time series (from 11 different subjects, Figure 5A). Figure 5B shows the distribution of the data in different time range before rehospitalization (from HF class).

After manual selection and labeling the five heartbeat ECG time series, we applied min-max normalization on each segment as follows:

$$X = \frac{\text{time serie} - \min(\text{time serie})}{\max(\text{time serie}) - \min(\text{time serie})}$$

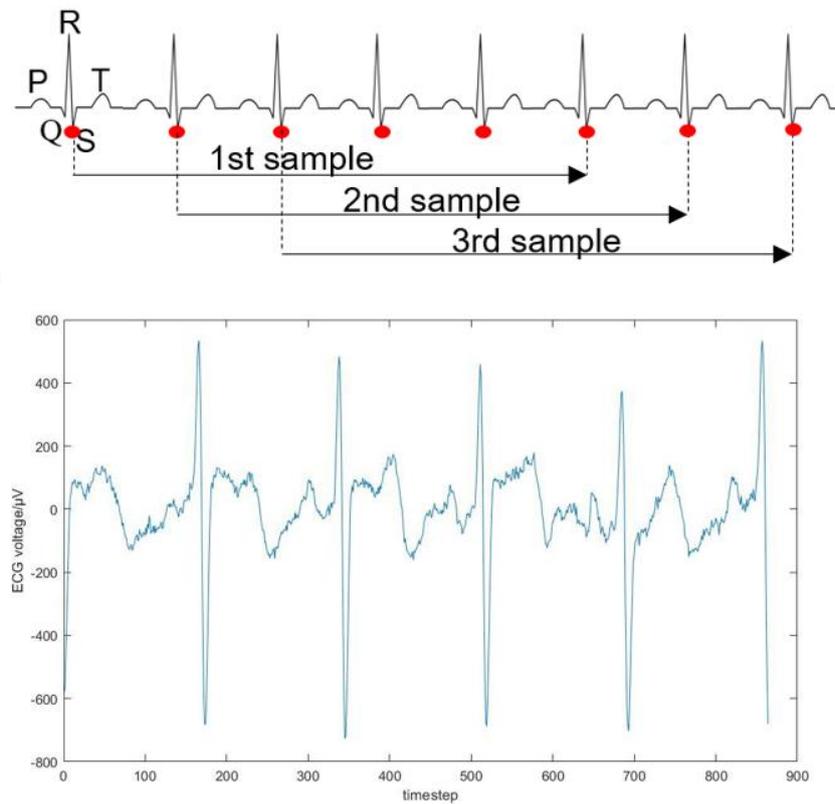


Figure 4: Remodeling of five minutes ECG time series into five heartbeat ECG time series. Upper: Schematic of raw ECG time series remodeling into five heartbeat ECG time series. Lower: example of resultant five heartbeat ECG time series.

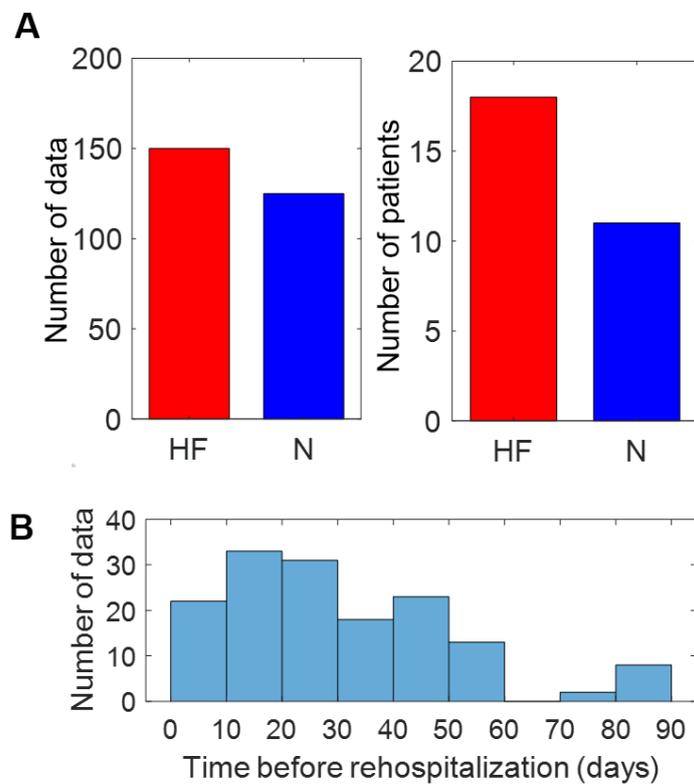


Figure 5: Distribution of data and source subjects among Class HF and N (A) and distribution of data in early dataset in accordance with distance in time to rehospitalization, which has been recognized as class HF (B).

2.3 Transformation from 1D time series into 2D images

We transferred normalized time series X to 2D greyscale images with Gramian Angular Summation Field (using $GASF()$ and $fit_transform()$ functions in $pyts.image$ python package). $GASF$ transformed 1D time series to 2D images by scaling time series into a normalized interval, followed by converting them into polar coordinate and calculating angular perspective among two different time points⁸⁴:

$$\tilde{x}_i = \frac{x_i - \max(X) + x_i - \min(X)}{\max(X) - \min(X)}$$

$$\phi = \arccos(\tilde{x}_i), -1 \leq \tilde{x}_i \leq 1, \tilde{x}_i \in \tilde{X}$$

$$r = \frac{t_i}{N}, t_i \in N$$

$$GASF_{ij} = \cos(\phi_i + \phi_j)$$

For the parameters in $GASF$, we used $overlapping = False$ to avoid Piece Aggregation Approximation (PAA) resizing for the time series and $scale = -1$ to set up boundaries for scaled time series \tilde{X} . After time series was transformed into gray-scaled images, we duplicated the images for three times to create RGB images as input for CNN. The resultant RGB images, as shown in were transferred and resized from ECG with varying temporal lengths into images with 224×224 in size by using $resize$ function from $cv2$ module, as the input for VGG19 (Figure 6).

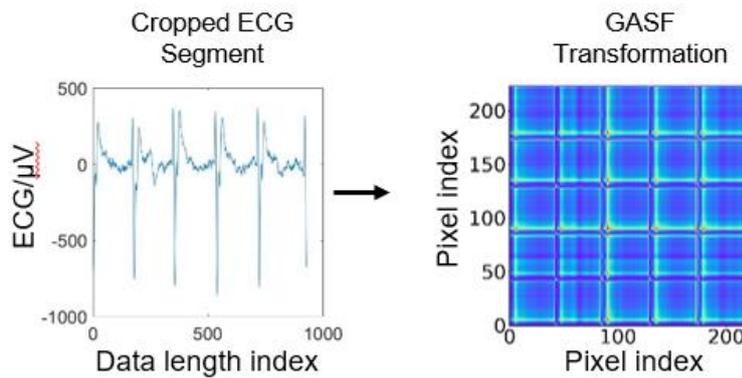


Figure 6: Transformation of ECG time series into 2D images by GASF, applied rainbow colormap for ease of visualization.

2.4 Feature extraction with VGG19 pre-trained neural network

The pre-trained VGG19 model took 2D images containing five beats ECG time series information

as input and produced features of these images as output. It contains 19 layers (16 convolutional layers and 3 fully-connected layers), and we extracted features after the first fully connected layer (FC1), which has a dimension of 4096 (Figure 7). The transfer learning model was weighted with ImageNet. After feature extraction, we applied PCA to reduce the dimension of raw features into 500 before applying classifier. We visualized VGG19-PCA500 features with t-Distributed Stochastic Neighbor Embedding (t-SNE)⁸¹.

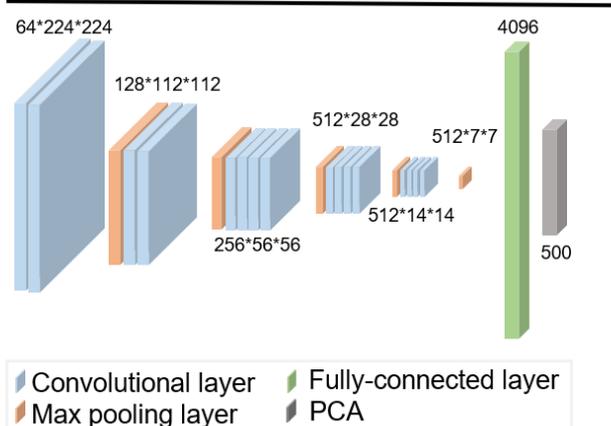
In order to reduce the negative effect of redundant features extracted from VGG19, in the initial attempt we used Principal Component Analysis (PCA)⁸⁵ dimensional reduction technique to eliminate the features that does not contains features that represent key patterns in ECG. These unwanted features can be regarded as noise, which has small influence in the illustration of morphological patterns of ECG curves, some of them might even bring negative influence. Here we used the entire dataset (4415 data) through image transformation, feature extraction of VGG19 and PCA dimensional reduction instead of directly processing through the earliest data. For one, the increase in the number of data create a “buffer” effect to PCA. PCA components act as linear combination of variables. The maximum number of components that 275 earliest data can have is 275 number of components. The reduction in that much number of dimensions might miss some important features. For another, using entire dataset provides higher learning ability for the model to distinguish the most important features from ECG time series. We trained a multilayer perceptron (MLP)⁸⁶ neural network using these VGG19-PCA features to build binary classifiers for the training data of Group HF and Group N with earliest 275 data only.

2.5 Multilayer perceptron neural network classification

We trained a multilayer perceptron (MLP) neural network using these VGG19-PCA features to build binary classifiers for the training data of Group HF and Group N. The classification used earliest data only. The whole training process was repeated 10 duplicated runs. Each run had random splitting of training (include validation) and testing sets with proportion of 0.8:0.2 respectively.

It consists of three neural network blocks. The first and second block consists of a FC layer (output dimensionality of 128 and 64, respectively, 0.01 for ‘regularizers.l2’), Batch Normalization (BN) layer, ReLU activation and Dropout layer (drop fraction of 0.5). The third block consists of FC layer with

Pre-trained convolution neural network VGG19 with PCA



Softmax activation.

Figure 7: Schematics of VGG19 model we used for feature extraction. Features were extracted from FC1 (green), followed by PCA dimensional reduction (grey).

Within the training set, we further used 10-fold sample-wise cross validation (to split training set and validation set). The classification model was fit on the training set, validated on 10 fold cross validation, with 64 epochs and class weight, which was calculated using `compute_class_weight()` function in `sklearn.utils` module. An early stop was used with minimum validation loss to prevent overfitting. The model was then used to predict label on test dataset.

In order to prove the hypothesis of existence of sample individuality, we further carried out a leave-one-subject-out cross validation. Each time, data from one individual patient was selected as the test set, while the data from the other 28 patients was selected as the training set. In order to further create randomness, we carried out 10-fold cross validation among the training set. The leave-one-subject-out cross validation was carried out for 29 times (because there were 29 patients in total). The test results from the same fold were concatenated together and regarded as the same run (for example, test results from fold 1 of all 29 leave-one-subject-out cross validation were regarded as run 1). In that way, the test result was weighted based on data source from different subject, meanwhile created a fluctuation due to randomness.

2.6 Standardization for sample individuality removal

In order to remove the sample individuality and clustering, we applied a simple, but successful standardization technique on the feature extracted from VGG19-PCA workflow. We used `StandardScaler` from `sklearn.preprocessing` module to carry out the standardization. In order to

remove the same distinctive pattern over VGG19-PCA features from the same subject/patient, standardization was applied to the data within the same patient in the direction of component-wise. The standardization was carried out on 29 batches, in each batch sample from one subject (patient) was fitted and transformed separately as follow (where x is the extracted feature from the same component, u and s are the mean and standard deviation of x , respectively):

$$z = \frac{x - u}{s}$$

2.7 Training classifier with support vector machine (SVM)

In order to implement the pipeline, we also applied another classifier, support vector machine (SVM)⁸⁷ instead of MLP classifier in the initial approach. In the main pipe line, the features extracted from VGG19-PCA-ST-UMAP structure were classified with Support Vector Machine classifier (SVM) after standardization and UMAP manifold learning. Due to the fact that in the initial attempt (no parameter tuning on UMAP, $n_neighbors$ and min_dist were default values of 15 and 0.1, respectively), visualization on the first two component of UMAP showed boundary between class HF and class N with circular shape (Figure 8). Then, using UMAP embedding with the optimum parameters, we trained the Support Vector Machine (SVM) classifier with a radial basis function kernel (most suitable to separate circular boundary).

2.8 Hyperparameter testing

Due to the fact that the specific effect of each parameter to classification performance remains unknown, we used a customized technique for parameter searching for UMAP parameters of $n_components$, $n_neighbors$, and min_dist , and SVM parameters of $gamma$ and c to get the combination of parameters with the optimum performance. The first step involves acquiring the optimum parameter for $n_neighbors$ and min_dist . We examined parameter searching for all combination of $n_neighbors$, in value of 2,3,5,7,10,15,20,25,30,40,50,75 and 100, and min_dist , in value of 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1. For simplicity, all calculations were carried out on UMAP with 25 number of components to get optimum $n_neighbors$ and min_dist parameters first. A grid searching was carried out for the combination of parameters $gamma$ and c on the magnitude of 0.001, 0.01, 0.1, 1, 10, 100, and 1000 (7*7 matrix with 49 results in total), with

GridSearchCV function from *sklearn.model_selection* module. Summation was carried out for each point with its surrounding neighbors to create a 5*5 sum matrix, as shown in Figure 9.

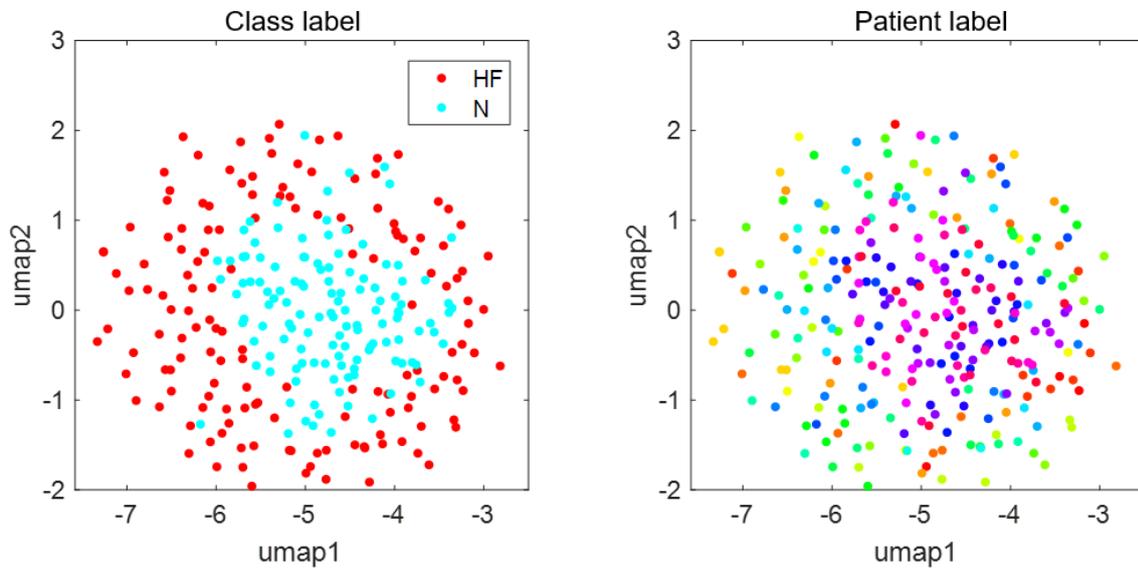


Figure 8: Visualization of initial attempt of the pipeline with default parameters in UMAP ($n_neighbors = 15$, $min_dist = 0.1$).

The further randomized search was carried out to the region that has the largest sum with *RandomizedSearchCV* function from the same module. During grid search and random search, cv was set with *PredefinedSplit* function from the same module, to create cross-validation, each time data from one patient was set as the test set while the others were used as the training set. The region was further divided into nine small blocks, each range from 0.5× to 5× for both its corresponding gamma and c values. Each block was iterated for 55 times (495 times for each combination of UMAP parameters), and the parameters with the best mean accuracy score were selected as the best parameter found for the specific combination of UMAP parameters. After the primary parameters for UMAP was decided, the same grid-randomized searching mechanism was applied to find out the optimum number of UMAP output components, and corresponding gamma and c values in SVM.

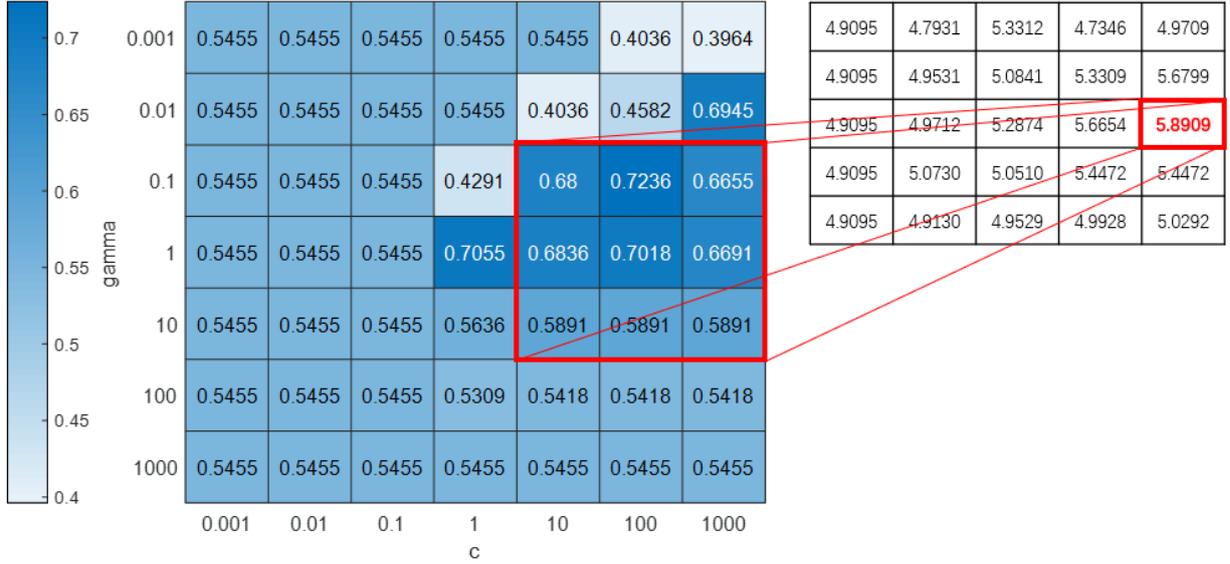


Figure 9: Example of initial step of parameter searching (in this case, maximum in summation field assigned boundaries of secondary random searching in nine small blocks with grid of γ in the value of 0.05, 0.5, 5, 50 and c in the value of 5, 50, 500, 5000).

2.9 Comparative methods as alternatives

In order to show the ability of GASF in interpretation of ECG patterns, we used Markov Transition Field and Recurrence Plot as comparison (using $MTF()$, $RP()$ and $fit_transform()$ functions in *pyts.image* python package)^{84,88}. As shown in Figure 10, similar as GASF, we duplicated the output gray-scaled images for three times as the RGB images input for CNN. As for MTF, a time series X is divided into Q quantile bins and each x_i is assigned into the corresponding q_j . By applying first-order Markov chain, an $Q \times Q$ sized adjacency matrix W with frequency w_{ij} (a point in q_j also fall into q_i) can be calculated⁸⁴. In this way, a time series can be converted to a 2D MTF matrix as follow:

$$MTF = \begin{bmatrix} w_{ij|x_1 \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_1 \in q_i, x_n \in q_j} \\ \vdots & \ddots & \vdots \\ w_{ij|x_n \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_n \in q_i, x_n \in q_j} \end{bmatrix}$$

As for RP, a time series is converted into its extracted trajectories, where m is the dimension of trajectories and τ is the time delay, and RP is calculated as the pairwise distance matrix between trajectories. We used 10% (default) as our percentage threshold ϵ :

$$\vec{x} = (x_i, x_{i+\tau}, \dots, x_{i+(m-1)\tau}), \forall i \in \{1, \dots, n - (m-1)\tau\}$$

$$RP_{ij} = \theta(\epsilon - \|\vec{x}_i - \vec{x}_j\|), \forall i, j \in \{1, \dots, n - (m-1)\tau\}$$

For both methods, the resultant RGB images were transferred and resized from ECG with varying

temporal lengths into images with 224×224 in size by using resize function from cv2 module, as the input of VGG19. As for these two methods, all the other steps remained constant.

We also tested the roles of the other components in our pipeline. To test the roles of SVM, we replaced SVM with MLP and trained the classifier (denoted as MLP). The previous pipeline associated with MLP classifier was further implemented with standardization and UMAP dimensional reduction as well.

We also experiment to quantitatively show how essential the feature standardization and UMAP steps are. NO_UMAP classifier contains all the other steps, but the data was classified with SVM right after standardization without UMAP dimensional reduction. And NO_STD classifier contains all the other steps, but the data has been processed through PCA dimension reduction and UMAP dimension reduction without feature standardization.

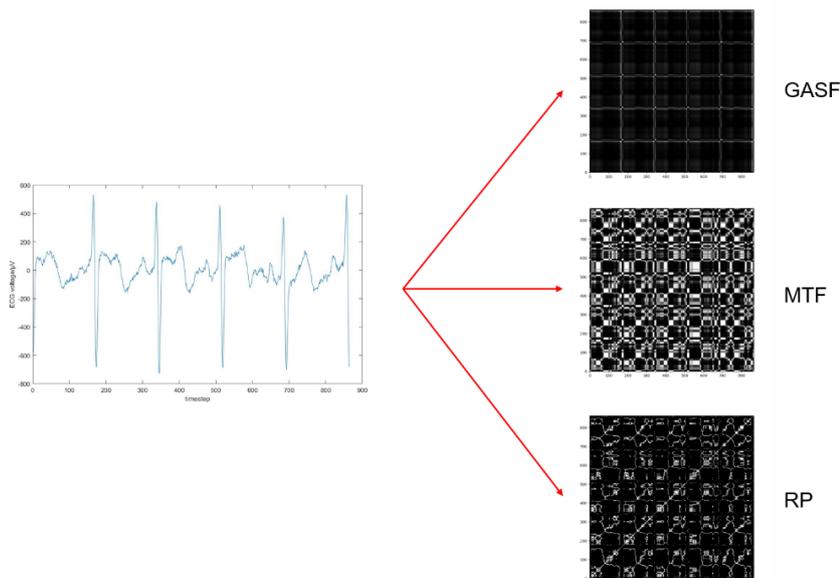


Figure 10: Schematics of image transformation process of a five heartbeat ECG time series into 2D images with different image transformation methods of GASF, MTF, and RP

To understand the roles of VGG19 feature extraction in our pipeline, we removed it from our pipeline and directly reduced the dimensionality of ECG time series using UMAP (UMAP_Only) and PCA (PCA_Only), followed by SVM training. For UMAP_Only, since the effect from different parameters and length on classification performance, we further applied hyperparameter searching on different parameters and length. As for PCA_Only, ECG was reduced to 115 dimensions that retained 95% variance. We also tested with another pre-trained framework (ResNet50)⁸⁹ for feature

extraction as comparison.

Chapter 3: Results

3.1 PCA dimensional reduction and 2D visualization with t-SNE on the reduced features

As shown in Figure 11, during the PCA dimensional reduction, although after the number of components increased to 50, the explained variance ratio was less than 0.001, we further found the cumulative explained variance reached 0.9945 at 500 number of components (compared with 0.9409 at 50 number of components). Therefore, PCA was applied to reduced the feature dimension from 4096 to 500, which retained more features that could be informative to classification and representation of ECG patterns.

We visualized the VGG19-PCA500 features through t-SNE on 2D scale. As shown in Figure 12, the plots exhibit an interesting phenomenon, that instead of clustering upon different classes (HF or N), the embedding result tend to form clusters based on the source of the data, which we referred as “patient clusters”. On the other hand, the difference in the clustering pattern of class HF and class N can be hardly observed. This phenomenon was later proved causing unwanted effects in classification.

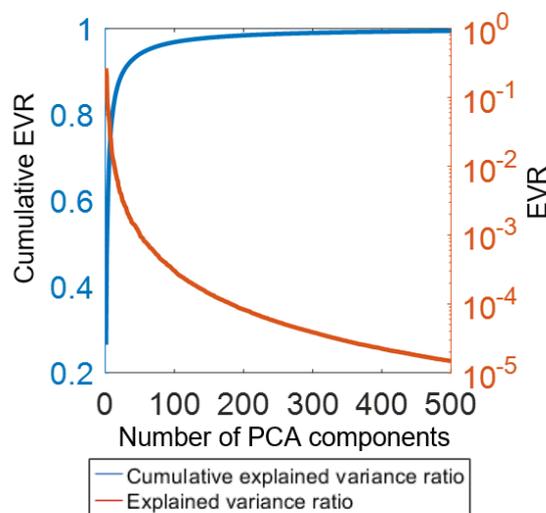


Figure 11: Cumulative explained variance ratio (EVR) and explained variance ratio plot with number of components from 1 to 500

3.2 Initial Classification based on MLP neural network classifier

The initial attempt of classification with MLP classifier showed promising result at first glance. The result shows that the average test accuracy of using pre-trained VGG19 to distinguish among different groups was $81.27 \pm 5.88\%$.

Although the initial classification over the pre-trained VGG19 features showed promising result, we hypothesize that due to the existence of patient cluster phenomenon, which can be observed during visualization, the result was affected that instead of predicting based on different classes, the model predicted based on different patients. In order to prove this, hypothesize, we further carried out a leave-one-subject-out cross validation. The corresponding average test accuracy showed huge reduction in performance to $50.84 \pm 4.16\%$. This result suggests that due to the effect of patient clusters, the VGG19 features themselves could incur a high degree of overfitting, and it is necessary to remove non-clinical patient bias from the VGG19-PCA features while retaining the heart failure-relevant information. Meanwhile, instead of predicting based on pattern of feature difference in class HF and class N, the model distinguished the distinctive pattern from each patient and predicted based on these differences.

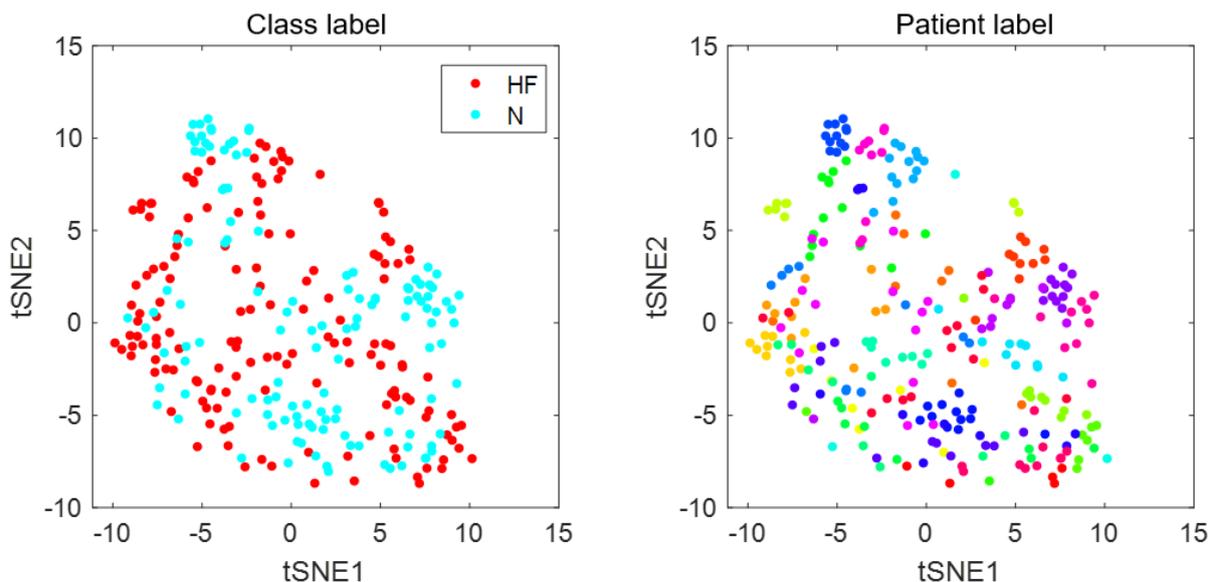


Figure 12: t-SNE visualization on 2D scale with class labels (left) and patient labels (right, each color represent one individual patient that provided ECG data)

The result for the primary classification and prediction provided two interpretations. On the one hand, this prediction cannot prove that whether the pipeline can actually be used to predict as the early indicator for recurrent heart failure worsening. But on the other hand, the distinction between different patients also represent the sensitivity of the model regarding distinctive features extracted from VGG19.

3.3 Balancing global and local data structures by UMAP

After applying the implementation on the pipeline, the sample individuality was eliminated (As shown in Figure 8). After applying UMAP dimensional reduction technique, the first two component of UMAP shows circular boundary between clusters of HF and N. We reasoned that UMAP balancing global and local data structures was able to preserve the clinically relevant global information while disregarded the local non-clinical patient individuality. The hyperparameter, $n_neighbors$ of UMAP controls this balancing act, and min_dist controls the tightness between points. The 2D visualization (Figure 13A) showed that UMAP with small $n_neighbors = 3$ preferentially preserved the local clusters while the large $n_neighbors (\geq 15)$ better conserved the global clinically relevant information. To quantitatively assess the performance for different combinations of UMAP hyperparameters, we applied grid searching to find out the optimum parameters in SVM. The classification was carried out with 13 different $n_neighbors$ parameters and 11 different min_dist parameters.

The classification result shows that the performance was better with large $n_neighbor (\geq 15)$ parameters and small $min_dist (\leq 0.2)$ (Figure 13B). This result suggests that balancing local and global data structures by UMAP is vital in predicting HF-related events while minimizing overfitting. The optimum performance was achieved when $n_neighbors = 30$ and $min_dist = 0$. In ease of calculation and time-efficiency, this combination of parameters was kept in use for alternatives of MTF, RP, ResNet, MLP and NO_ST.

3.4 Implemented new pipeline with standardization, manifold learning of UMAP and SVM classifier

The parameter searching in UMAP showed that the combination of combination of $gamma = 30$ and $c = 0$ preserved the best balance of global and local structure balance, thus create the optimum

performance in classification. After applying UMAP, HF-related classes form the clear circular shape boundary (Figure 14A, left) while the small patient clusters had been eliminated (Figure 14A, right). The patient-wise clustering of ECG signals was no longer observed. The data from the same subject now scatters all around within the cluster. In fact, the new feature extraction pipeline segregated the data only into the HF-related classes into the center of the plot as one large cluster.

We applied the same classification technique previously used to evaluate the model. As for testing set, we applied 29 leave-one-subject(patient)-out classification, and for model selection and validation we applied 10-fold cross-validation. Due to the fact that the length of number of components from UMAP stays unknown, we further applied another grid searching from 1 component to 100 components. By applying the customized parameter search technique for hyper-parameter tuning, we found optimum parameters at components of UMAP = 35, $n_neighbors = 30$, $min_dist = 0.0$, $gamma = 0.45$, and $C = 20.55$, with average test accuracy of $91.48\% \pm 0.31\%$ and area under receiver operating characteristics (AuROC) of 0.9470 (Figure 14C, D). This result demonstrated that the model has the feasibility of the early prediction of heart failure worsening based on ECG signals.

3.5 Roles of each computational step in our pipeline

The detailed classification results and corresponding optimum parameter combination being used to achieve the classification result is shown in Table 1. As for UMAP_Only, optimum parameters were found in parameter search at components of UMAP = 21, $n_neighbors = 100$, $min_dist = 0.0$, $gamma = 0.33$, and $C = 38.88$. The average weighted accuracy was $85.56\% \pm 0.18\%$, which is significantly lower than the performance of the main workflow (Figure 14C, Table 1). For PCA_Only, the classification accuracy was $45.49\% \pm 0.95\%$ (Figure 14C, Table 1). Furthermore, VGG19 features were shown to provide richer features to the downstream UMAP, resulting in a significant boost of the performance (denoted as Main approach in Figure 14C).

As for alternative in associated with feature extraction of ResNet50 (RESNET), similar to VGG19, the features also showed segregation of class clusters in 2D UMAP plot (Figure 14B). The patient individuality has been eliminated. However, the boundary between the cluster of Class HF with Class N is not as distinct as the feature processed with VGG19. More data scatters around and invade into the boundary, resulting in a region that contains both classes of data. With $n_neighbors$ of 30 and min_dist of 0.0 for UMAP, the optimum average weighted test accuracy was $84.62\% \pm 0.34\%$ (number

of UMAP components = 8, gamma \approx 0.44 and $c \approx$ 20.55).

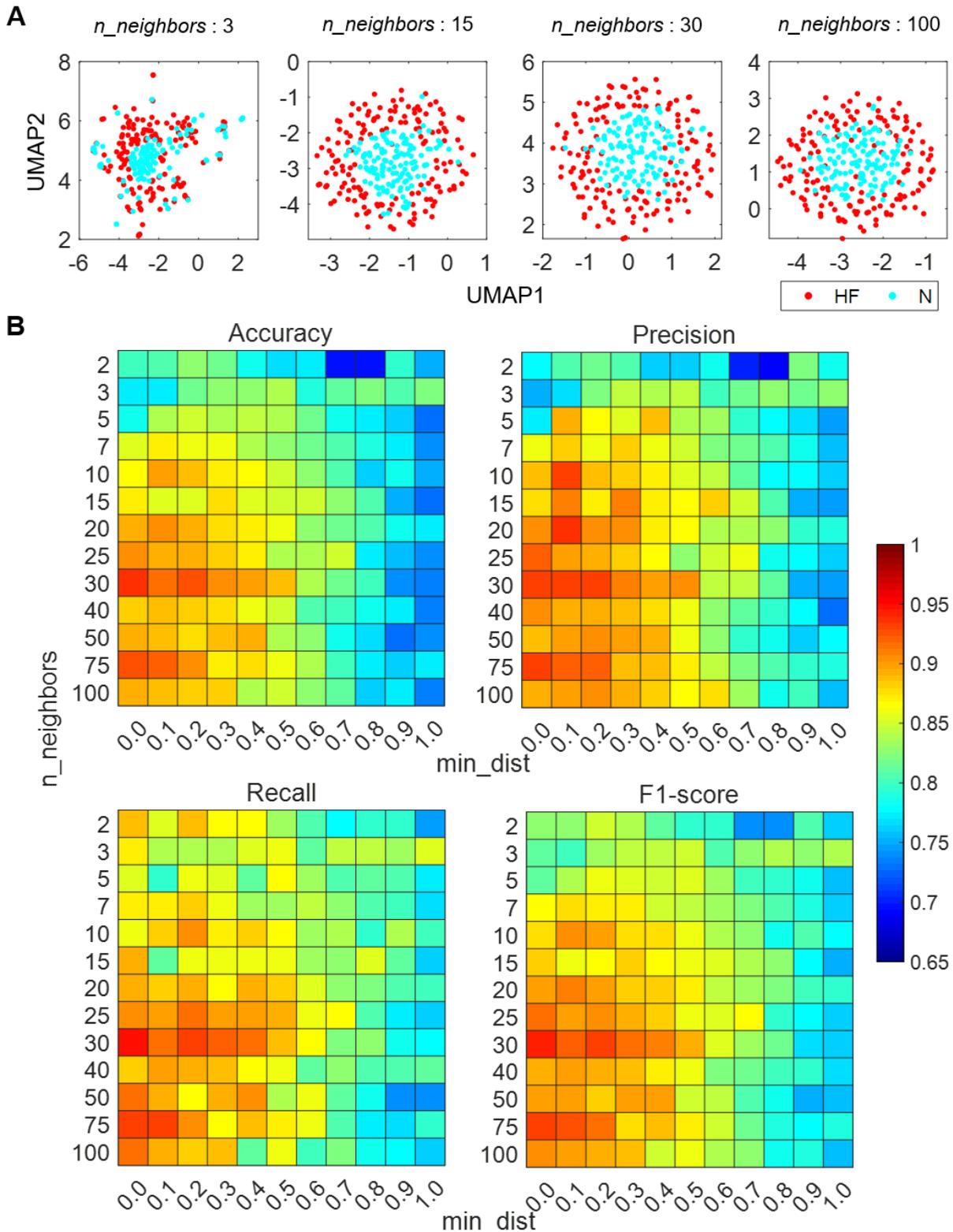


Figure 13: Balancing global and local data structure by UMAP. (A) 2 dimensional UMAP visualization with $min_dist = 0.0$ and $n_neighbors$ parameters set to 3, 15, 30 and 100. (B) Randomized search weighted scoring results with varying $n_neighbors$ and min_dist hyper-parameters to balance global and local structures of the features based on main approach dataset with 35 $n_components$

	NoC	NN	MD	Gamma	c	Accuracy	Precision	Recall	F1-score
Main approach	35	30	0.0	0.45	20.55	0.9213±0.0028	0.9293±0.0021	0.9261±0.0048	0.9277±0.0027
ResNet	8			0.13	42.13	0.8141±0.0034	0.8165±0.0018	0.8503±0.0062	0.8330±0.0035
MLP	2					0.8218±0.0270	0.7861±0.0343	0.9280±0.0172	0.8507±0.0201
NO_ST	19			16.06	5.21	0.5982±0.0088	0.6005±0.0075	0.7876±0.0057	0.6814±0.0045
MTF	39			0.04	1291.85	0.7850±0.0045	0.8151±0.0034	0.7836±0.0064	0.7990±0.0046
RP	41			0.30	17.12	0.8021±0.0042	0.8346±0.0024	0.7945±0.0072	0.8141±0.0045
No_UMAP						0.0005	303.22	0.4797±0.0039	0.4770±0.0045
UMAP_Only	21	100	0	0.33	38.88	0.8556±0.0018	0.8841±0.0012	0.8463±0.0036	0.8648±0.0019
PCA_Only				0.001	174.63	0.4549±0.0097	0.4567±0.0095	0.4755±0.0126	0.4659±0.0109

Table 1: Optimum classification performance main approach and alternative pipelines achieved and corresponding parameters

As for alternatives method MLP, we found optimum performance when number of UMAP components = 2, average weighted test accuracy was 82.18%±2.70%. As a comparison, our method with VGG19 and SVM has much higher performance in accuracy, precision, recall and F1 score (Figure 14C) than MLP case.

As for NO_UMAP, we applied t-SNE visualization with the features before classification, we were not able to see any segregation of data, either among HF class labels or patients label (Figure 15). Although the standardization step was proven to be useful to remove patient individuality, we were not able to successfully classify features into different classes without UMAP (classification accuracy: 47.97%±0.39%). As for NO_STD, the 2D UMAP plot showed segregation among patient clusters (Figure 16, right), while the large clusters of the HF classes could not be observed (Figure 16, left). The best performance was found with the test accuracy of 59.82%±0.88% with optimum parameters of UMAP components = 19, gamma ≈ 16.06, and c ≈ 5.21.

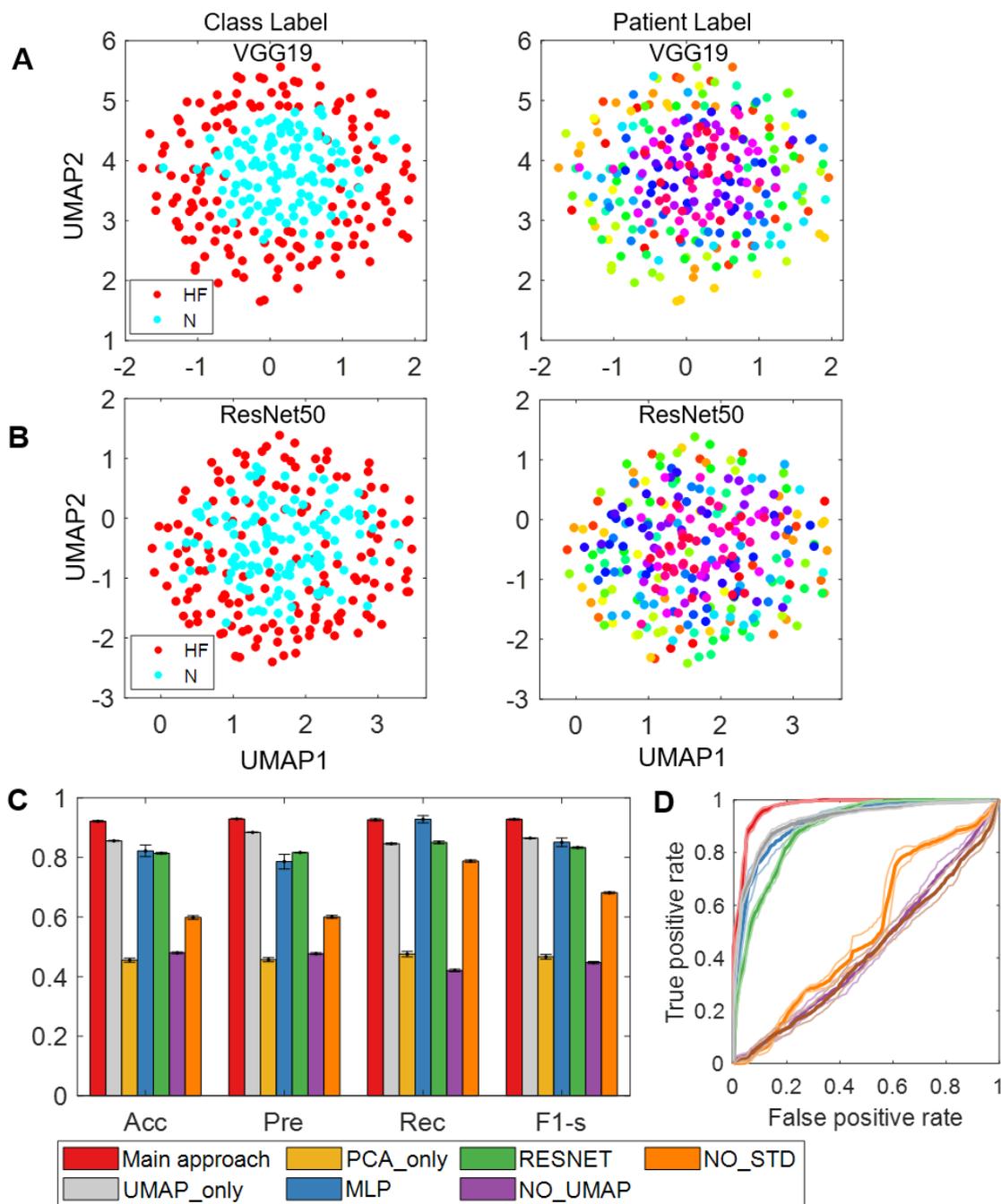


Figure 14: Classification performance on feature extracted from early prediction of heart failure worsening ECG data. (A-B) 2D UMAP visualization of the standardized features from the methodology associated with raw feature extraction with VGG19 (A) and ResNet50 (B) models. Left of (A-B): Labeled with different classes (HF/N). Right of (A-B): Each color represents one individual patients. (C) Classification performance among the main approach and alternatives (D) Receiver Operating Characteristic among the classification result from main approach and alternatives (red: mean ROC curve of main approach, mean Auc = 0.9716; grey: UMAP_only, mean Auc = 0.9203; claret: PCA_only, mean Auc = 0.4149; blue: MLP, mean Auc = 0.9196; green: RESNET, mean Auc = 0.8879; purple: NO_UMAP, mean Auc = 0.5038; orange: NO_STD, mean Auc = 0.4295; transparent thin curves: 95% confidence interval among 10 folds).

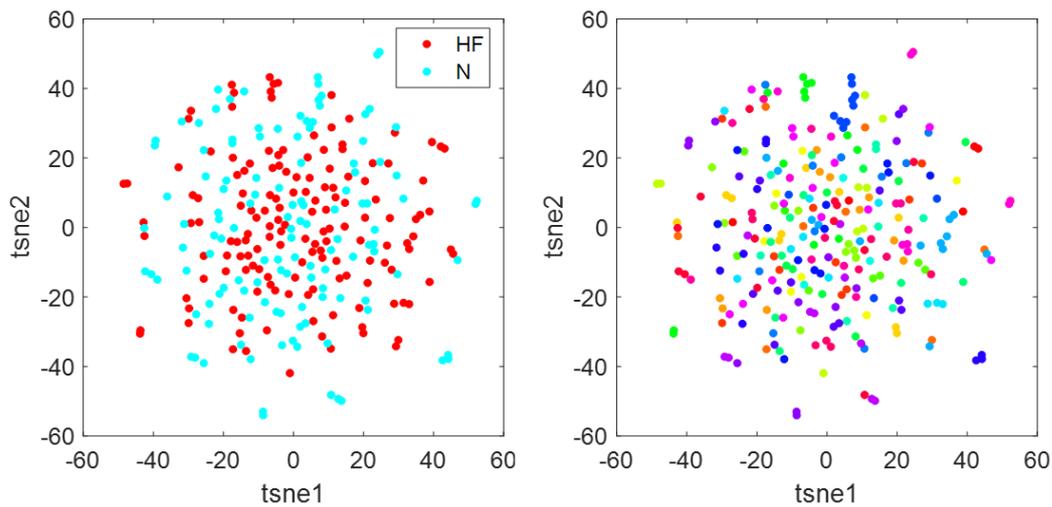


Figure 15: 2D t-sne visualization on No_UMAP (Left: class label. Right: patient label, each color represents one individual patient)

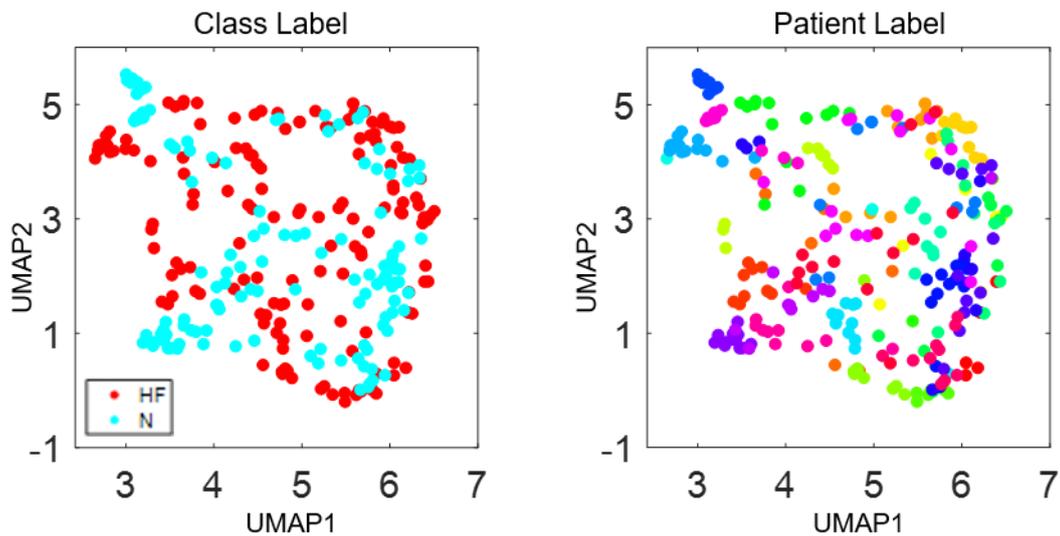


Figure 16: 2D UMAP visualization on No_STD (Left: class label. Right: patient label, each color represents one individual patient)

We also tested the importance of GASF transformation in our pipeline. With Markov Transition Field (MTF), and Recurrence Plot (RP), (see Methods section for detail), we applied the same structure of our main approach. By visualizing through the first two components of UMAP, the class boundary between group HF and group N was not as obvious as that from GASF (Figure 17A). However, the segregation of data forming small patient clusters was no longer present in all plots, indicating the methodology was valid in eliminating sample individuality with other image transformation methods as well. Significantly lower classification performances and AUC also suggest the better performance of GASF (Figure 17B,C, Table 1).

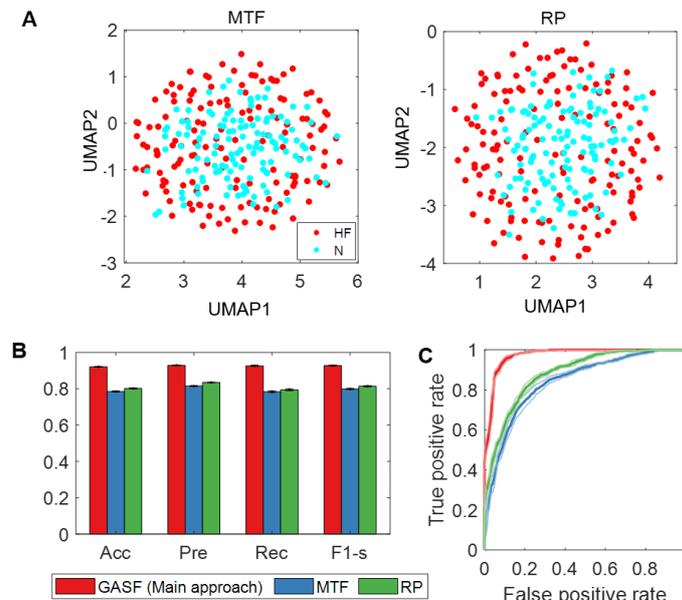


Figure 17: Classification performance comparison among different image transformation methods. (A) 2D UMAP visualization of VGG19-PCA features using MTF (left) and RP (right). (B) Classification performance among different transformation methods. (C) Receiver Operating Characteristic curves (mean GASF = 0.9716; mean MTF = 0.8380; mean RP = 0.8794; transparent thin curves: 95% confidence interval among 10 folds).

Chapter 4: Conclusion and outlook

The initial approach in this study started by testing the sensitivity of the model with the entire dataset of SENTINEL-HF real-life dataset. The approach aimed at showing classification robustness of the model distinguishing time-series data with different patterns. The raw inputs of the model were images that transferred from five heartbeat ECG time series by Gramian Angular Summation Field (GASF). The extraction of features was carried out by pre-trained model VGG19 weighted with ImageNet, while the classification was carried out by a multilayer perceptron (MLP) classifier. Although the initial result shows promising in classification performance, it also shows sample individuality. The sample individuality found in the initial approach created bias during classification. Instead of distinguishing the common pattern among different class (heart failure or normal), the classifier recognized the same pattern among the same subject/patient. Therefore, a standardization was applied after the step of PCA dimensional reduction (which aims at removing features that is not useful, or can cause negative effect), followed by a UMAP dimensional reduction (which aims at further create clusters among same classes with better balanced local and global structure). We also found Support Vector Machine (SVM) outperforms than MLP in this scenario of prediction on early recurrent heart

failure indication.

The comparison between the main approach and alternatives proved that all the components in the workflow of VGG19-PCA-ST-UMAP-SVM are essential to carry out the most accurate prediction on early recurrent heart failure indication with the real-life dataset from SENTINEL-HF study. Although feature extraction from ResNet50 model also exhibited promising result, VGG19 outperforms it with classification performance. Meanwhile as for classifier, SVM outperforms than MLP with less calculation time and better prediction ability and distinguishing early heart failure indication and normal cases. Furthermore, switching to other image transformation technique (MTF and RP) showed lower classification performance and AUC, suggest GASF in the ability to interpret and emphasize the pattern in ECG which indicates the likelihood of recurrent heart failure. Although the grey scaled image transformed by MTF and RP showed more information than the image transformed by GASF (Figure 10), the redundancy in the information acted as noise and caused negative effect in feature extraction. Most of all, comparing with the result from NO_ST, the success in applying standardization step to remove sample individuality makes the classification focused on the distinctive features that indicates the early sign of heart failure.

With the increase in $n_neighbors$ parameter, UMAP tend to present more of the global structure of the data. However, the classification performance at $n_neighbors = 30$ was much better than when $n_neighbors$ was even larger. The hyperparameter study suggests that although the preservation of global structure is one of the advantages of UMAP, the classification works best when global structure and local structure reaches a balance. When UMAP was not applied to the features, as shown in Figure 14, although standardization removed sample individuality (no more subject clusters), the segregation of class clusters cannot be successfully carried out without processing with UMAP dimensional reduction. These results suggest that the manifold learning by UMAP played a critical role in our pipeline to re-gather and present the most important feature in low-dimensional structure after standardization. It allowed us to achieve a good performance, even directly using ECG time series as the input. Therefore, we find out that parameters that are influential in global structure and local structure preservation, for example $n_neighbors$ and min_dist , and output dimensionality $n_components$, should be examined before applying the post-processed features with optimum UMAP parameters on classifiers.

The previous deep learning studies were focused on detecting heart failure events, which limits the ability of the early prediction^{54,90}. Here, we showed the feasibility of predicting the precursor of heart failure (heart failure-worsening), which enables early alerts and interventions of heart failure worsening. The performance of studies in the relationship between heart failure worsening and ECG is highly restricted to the proper positioning of electrodes. Unlike the data in majority public databases that has been widely accepted and used in other studies, which was recorded by professionals with highly precise equipment, raw data in this study was recorded by patients themselves at home with close-fitting vest which carries four electrodes. This study provides a new approach in classifying ECG across different channels and improper positioning of electrodes, and a solution in detecting heart failure worsening with portable health care devices.

The result from this study is promising. However, it also has few limitations. First of all, the early dataset only contains 275 data from 29 patients in total. In order to make sure VGG19-PCA500 feature extraction model captures the most representative features, the early dataset was required to combine with later data to achieve better performance. The limitation of sample size also limits the scale of the study in predicting whether the data exhibiting early sign of recurrent heart failure. With a larger sample and subject size, the prediction can be carried out more quantitatively. For example, the likelihood of a person is going to have recurrent heart failure after discharged from hospital. And if heart failure is likely to recur, how far is it from the day of discharge (multi-classes classification instead of binary classification). However, the result in this study actually provided a future path that, it is possible to mix raw ECG data extracted from different studies once they are in accordance with the same standard (for example, contains same amount of channels and location/positioning of the electrodes), so that sample size can be enlarged easily. Another limitation is that this study is only applicable to single channel ECG, while the effect and adjustment of this approach with multi-channel ECG. In hospitals, when cardiologists measure ECG on patients, the maximum number of leads is usually 12, the ECG data being applied in clinics is often in 12 channel form. Therefore, one of our future work related to this project is implementing the methodology to use conventional multiple ECG channels, which is expected to produce better results.

From the clinical perspective, the existence of this methodology provided a potential solution and evaluation method for patient recovery after initial treatment of heart failure (during discharge). With proper, routine measurement of ECG, the status of patient and risk for heart failure worsening can be

measured remotely when patients are at home, instead of visiting hospitals for weekly re-examination, which is more convenient and efficient. However, before applying the methodology to clinical practices, there is one aspect that should be taken into consideration. The clinical practices mean the methodology works for majority of people, therefore, we must prove that the methodology can still have good performance with larger dataset and sample size.

Analyzing time series data with the transformation into image-based data provided a practical approach in transfer learning. Transfer learning with pre-trained model provides the result with high accuracy but short computation time and increases the feasibility of machine learning applications in the medical field. To some extent, the application of this approach is not only limited to the prediction of Other possible applications of this technique include analysis of the Electroencephalogram (EEG) in epilepsy detection⁹¹, Electroretinogram (ERG) in congenital stationary night blindness diagnosis⁹², and the classification/clustering of dynamic cellular phenotypes⁹³. However, depending on different datasets and applications, the detailed methodology and components might vary. When the raw data is acquired, depending on the data type, different pre-processing method might be applied for adoptability. If the raw data is in image form, it can be directly used as input for our machine learning models. However, if the data type is time series, it should be transferred into images first (similar to our approach of using GASF). Then the size of dataset should be taken in to consideration for feature extraction model selection. Similar to this study, transfer learning might be a good choice for small dataset to avoid overfitting. However, different models, for example conventional CNN and LSTM might also be tested at the beginning and compare the performance to select the most suitable model for specific dataset. The feature extraction step is followed by an optional initial dimensional reduction step to remove undesired redundancy in dimensions and simplify dataset. After visualization on 2D structure (UMAP, t-SNE or PCA), depending on clustering situation (whether sample individuality exists), a standardization step can be added before manifold learning and classification. For example, with samples collected from very accurate and uniform equipment from the same subject, the sample individuality might not be significant, therefore the step of standardization might not be required. Because when applying UMAP, the condition of local and global structure is often significantly different depending on specific dataset, in the manifold learning (UMAP) step a parameter searching step should be included (in associated with classification method) to find out the optimum balance (trade-off) between local and global structures in their influence at classification performance. The detailed

flow chart is shown in Figure 18 below.

In conclusion, in this thesis, we developed a time series transfer learning pipeline, which can effectively classify patients' ECG time series data into important clinical conditions without detecting known abnormal ECG rhythms. With the pre-trained convolutional neural network on its application to heart failure screening, high accuracy was achieved with the inputs that consist of transferred 2D images from 1D time series, followed by the feature standardization and manifold learning. For the prediction of early sign of recurrent heart failure worsening, the feature standardization and UMAP manifold learning in our pipeline are able to effectively remove the nonclinical patient individuality and keep the information highly relevant to the heart failure worsening. Furthermore, with the structure of combining the standardized VGG19 features with UMAP, we conclude that our new approach can carry out high-performance classification with data that contains individuality from sources and sample bias.

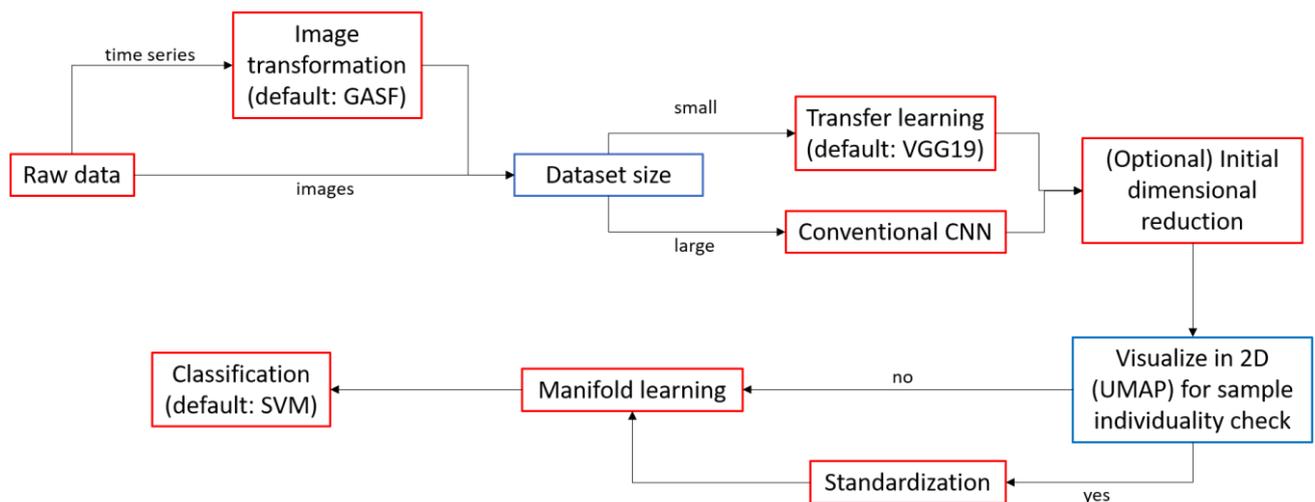


Figure 18: Suggested workflow of the application of this methodology on other dataset and studies

References

- 1 Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of research and development* **3**, 210-229 (1959).
- 2 O'Shea, T. J., Corgan, J. & Clancy, T. C. in *International Conference on Engineering Applications of Neural Networks*. 213-226 (Springer).
- 3 Myers, M. J. Voice recognition software and a hand-held translation machine for second-language learning. *Computer Assisted Language Learning* **13**, 29-41 (2000).
- 4 Hsu, R.-L., Abdel-Mottaleb, M. & Jain, A. K. Face detection in color images. *IEEE transactions on pattern analysis and machine intelligence* **24**, 696-706 (2002).
- 5 Ayata, D., Yaslan, Y. & Kamasak, M. E. Emotion based music recommendation system using wearable physiological sensors. *IEEE transactions on consumer electronics* **64**, 196-203 (2018).
- 6 Chen, H.-C. & Chen, A. L. P. in *Proceedings of the tenth international conference on Information and knowledge management*. 231-238.
- 7 Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484-489, doi:10.1038/nature16961 (2016).
- 8 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444, doi:10.1038/nature14539 (2015).
- 9 Chen, J. X. The evolution of computing: AlphaGo. *Computing in Science & Engineering* **18**, 4-7 (2016).
- 10 Borowiec, S. AlphaGo seals 4-1 victory over Go grandmaster Lee Sedol. *The Guardian* **15** (2016).
- 11 Lee, K.-F. *AI superpowers: China, Silicon Valley, and the new world order*. (Houghton Mifflin Harcourt, 2018).
- 12 Berner, C. *et al.* Dota 2 with Large Scale Deep Reinforcement Learning. *arXiv preprint arXiv:1912.06680* (2019).
- 13 Ren, X., Wang, Y., Chen, L., Zhang, X.-S. & Jin, Q. ellipsoidFN: a tool for identifying a heterogeneous set of cancer biomarkers based on gene expressions. *Nucleic acids research* **41**, e53-e53 (2013).
- 14 Hanka, R., Harte, T. P., Dixon, A. K., Lomas, D. J. & Britton, P. D. Neural networks in the interpretation of contrast-enhanced magnetic resonance images of the breast. *CURRENT PERSPECTIVES IN HEALTHCARE COMPUTING*, 275-283 (1996).
- 15 Stausberg, J. & Person, M. A process model of diagnostic reasoning in medicine. *International Journal of Medical Informatics* **54**, 9-23 (1999).
- 16 Zupan, B., Halter, J. A. & Bohanec, M. Qualitative model approach to computer assisted reasoning in physiology. *Proceedings of Intelligent Data Analysis in Medicine and Pharmacology-IDAMAP98* (1998).
- 17 Li, Y. & Cui, W. Identifying the mislabeled training samples of ECG signals using machine learning. *Biomedical Signal Processing and Control* **47**, 168-176 (2019).
- 18 LeCun, Y. & Bengio, Y. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* **3361**, 1995 (1995).
- 19 Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097-1105 (2012).
- 20 Sun, W., Zheng, B. & Qian, W. Computer aided lung cancer diagnosis with deep learning algorithms. *Medical imaging 2016: computer-aided diagnosis* **9785**, 97850Z (2016).
- 21 Kirienko, M. *et al.* Convolutional neural networks promising in lung cancer T-parameter assessment on baseline FDG-PET/CT. *Contrast Media & Molecular Imaging* **2018** (2018).
- 22 Allen, L. A. & O'Connor, C. M. Management of acute decompensated heart failure. *Cmaj* **176**, 797-805 (2007).
- 23 Joseph, S. M., Cedars, A. M., Ewald, G. A., Geltman, E. M. & Mann, D. L. Acute decompensated heart failure: contemporary medical management. *Texas Heart Institute Journal* **36**, 510 (2009).

-
- 24 Fang, J., Mensah, G. A., Croft, J. B. & Keenan, N. L. Heart failure-related hospitalization in the US, 1979 to 2004. *Journal of the American College of Cardiology* **52**, 428-434 (2008).
- 25 Garde, A. H., Hansen, Å. M., Holtermann, A., Gyntelberg, F. & Suadicani, P. Sleep duration and ischemic heart disease and all-cause mortality: prospective cohort study on effects of tranquilizers/hypnotics and perceived stress. *Scandinavian journal of work, environment & health*, 550-558 (2013).
- 26 Stansfeld, S. A. & Marmot, M. G. *Stress and the heart: Psychosocial pathways to coronary heart disease*. (BMJ books, 2002).
- 27 Freedman, D. S., Khan, L. K., Dietz, W. H., Srinivasan, S. R. & Berenson, G. S. Relationship of childhood obesity to coronary heart disease risk factors in adulthood: the Bogalusa Heart Study. *Pediatrics* **108**, 712-718 (2001).
- 28 Centers for Disease, C. & Prevention. Sodium intake among adults-United States, 2005-2006. *MMWR. Morbidity and mortality weekly report* **59**, 746 (2010).
- 29 Writing Group to Review New, E. *et al.* 2009 focused update: ACCF/AHA guidelines for the diagnosis and management of heart failure in adults: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines: developed in collaboration with the International Society for Heart and Lung Transplantation. *Circulation* **119**, 1977-2016 (2009).
- 30 Lala, A. *et al.* Relief and recurrence of congestion during and after hospitalization for acute heart failure: insights from Diuretic Optimization Strategy Evaluation in Acute Decompensated Heart Failure (DOSE-AHF) and Cardiorenal Rescue Study in Acute Decompensated Heart Failure (CARESS-HF). *Circulation: Heart Failure* **8**, 741-748 (2015).
- 31 Ahmed, A. *et al.* Incident heart failure hospitalization and subsequent mortality in chronic heart failure: a propensity-matched study. *Journal of cardiac failure* **14**, 211-218 (2008).
- 32 Solomon, S. D. *et al.* Influence of nonfatal hospitalization for heart failure on subsequent mortality in patients with chronic heart failure. *Circulation* **116**, 1482-1487 (2007).
- 33 De Caterina, R. *et al.* ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2008. *Eur J Heart Fail* **10**, 933-989 (2008).
- 34 Thygesen, K. *et al.* Third universal definition of myocardial infarction. *Circulation* **126**, 2020-2035, doi:10.1161/CIR.0b013e31826e1058 (2012).
- 35 O'Neal, W. T. *et al.* Electrocardiographic Predictors of Heart Failure With Reduced Versus Preserved Ejection Fraction: The Multi-Ethnic Study of Atherosclerosis. *J Am Heart Assoc* **6**, doi:10.1161/JAHA.117.006023 (2017).
- 36 Lyon, A., Mincholé, A., Martínez, J. P., Laguna, P. & Rodríguez, B. Computational techniques for ECG analysis and interpretation in light of their contribution to medical advances. *J R Soc Interface* **15**, doi:10.1098/rsif.2017.0821 (2018).
- 37 Obermeyer, Z. & Emanuel, E. J. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* **375**, 1216-1219, doi:10.1056/NEJMp1606181 (2016).
- 38 Wang, Z. & Oates, T. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015).
- 39 Acharya, U. R. *et al.* Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. *Information Sciences* **415**, 190-198 (2017).
- 40 Tan, J. H. *et al.* Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. *Comput Biol Med* **94**, 19-26, doi:10.1016/j.combiomed.2017.12.023 (2018).
- 41 Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S. & Acharya, U. R. Deep learning for healthcare applications based on physiological signals: A review. *Comput Methods Programs Biomed* **161**, 1-13,

-
- doi:10.1016/j.cmpb.2018.04.005 (2018).
- 42 Candelieri, A. *et al.* Early detection of decompensation conditions in heart failure patients by knowledge discovery: the HEARTFAID approaches. *2008 Computers in Cardiology*, 893-896 (2008).
- 43 Sevakula, R. K. *et al.* State-of-the-Art Machine Learning Techniques Aiming to Improve Patient Outcomes Pertaining to the Cardiovascular System. *Journal of the American Heart Association* **9**, e013924 (2020).
- 44 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. *Computational and structural biotechnology journal* **15**, 26-47 (2017).
- 45 Moody, G. B. & Mark, R. G. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine* **20**, 45-50 (2001).
- 46 Baim, D. S. *et al.* Survival of patients with severe congestive heart failure treated with oral milrinone. *Journal of the American College of Cardiology* **7**, 661-670 (1986).
- 47 Hamilton, P. S. & Tompkins, W. J. Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. *IEEE transactions on biomedical engineering*, 1157-1165 (1986).
- 48 Saini, I., Singh, D. & Khosla, A. QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases. *Journal of advanced research* **4**, 331-344 (2013).
- 49 Melo, S. L., Caloba, L. P. & Nadal, J. Arrhythmia analysis using artificial neural network and decimated electrocardiographic data. *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)* **27**, 73-76 (2000).
- 50 Masetic, Z. & Subasi, A. Congestive heart failure detection using random forest classifier. *Computer methods and programs in biomedicine* **130**, 54-64 (2016).
- 51 Acharya, U. R. *et al.* Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals. *Applied Intelligence* **49**, 16-27 (2019).
- 52 Hannun, A. Y. *et al.* Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine* **25**, 65 (2019).
- 53 Shanmugam, D., Blalock, D. & Guttag, J. Multiple Instance Learning for ECG Risk Stratification. *arXiv preprint arXiv:1812.00475* (2018).
- 54 Porumb, M., Iadanza, E., Massaro, S. & Pecchia, L. A convolutional neural network approach to detect congestive heart failure. *Biomedical Signal Processing and Control* **55**, 101597 (2020).
- 55 Attia, Z. I. *et al.* Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nature medicine* **25**, 70-74 (2019).
- 56 Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning*. (MIT press, 2016).
- 57 Zhang, W. in *annual conference of the Japan Society of Applied Physics*.
- 58 LeCun, Y. *et al.* Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective* **261**, 276 (1995).
- 59 Hijazi, S., Kumar, R. & Rowen, C. Using convolutional neural networks for image recognition. *Cadence Design Systems Inc.: San Jose, CA, USA*, 1-12 (2015).
- 60 Fu, J., Zheng, H. & Mei, T. in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4438-4446.
- 61 Hu, B., Lu, Z., Li, H. & Chen, Q. in *Advances in neural information processing systems*. 2042-2050.
- 62 Zhu, Y. *et al.* in *Twenty-Fifth AAAI Conference on Artificial Intelligence*.
- 63 Deng, J. *et al.* 248-255 (Ieee).
- 64 Kim, S.-J. *et al.* Deep transfer learning-based hologram classification for molecular diagnostics. *Scientific reports* **8**, 1-12 (2018).
- 65 Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. 3320-3328.
- 66 Sharif Razavian, A., Azizpour, H., Sullivan, J. & Carlsson, S. 806-813.
- 67 Donahue, J. *et al.* A deep convolutional activation feature for generic visual recognition. *arXiv preprint*

-
- 68 *arXiv:1310.1531* **1** (2013).
- 69 Oquab, M., Bottou, L., Laptev, I. & Sivic, J. 1717-1724.
- 70 Zhao, W. in *AIP Conference Proceedings*. 1 edn 020018 (AIP Publishing LLC).
- 71 Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- 72 Dovancescu, S. *et al.* Detecting heart failure decompensation by measuring transthoracic bioimpedance in the outpatient setting: rationale and design of the SENTINEL-HF study. *JMIR research protocols* **4**, e121 (2015).
- 73 Olszewski, R. T. Generalized feature extraction for structural pattern recognition in time-series data. (CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 2001).
- 74 Talwalkar, A., Kumar, S. & Rowley, H. in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 1-8 (IEEE).
- 75 Cunningham, J. P. & Ghahramani, Z. Linear dimensionality reduction: Survey, insights, and generalizations. *The Journal of Machine Learning Research* **16**, 2859-2900 (2015).
- 76 Sumithra, V. S. & Surendran, S. A review of various linear and non linear dimensionality reduction techniques. *Int. J. Comput. Sci. Inf. Technol.* **6**, 2354-2360 (2015).
- 77 Tenenbaum, J. B., De Silva, V. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science* **290**, 2319-2323 (2000).
- 78 Roweis, S. T. & Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *science* **290**, 2323-2326 (2000).
- 79 Zhang, Z. & Wang, J. MLL: Modified locally linear embedding using multiple weights. *Advances in neural information processing systems.*, 1593-1600 (2007).
- 80 Donoho, D. L. & Grimes, C. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences* **100**, 5591-5596 (2003).
- 81 Belkin, M. & Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* **15**, 1373-1396 (2003).
- 82 Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9**, 2579-2605 (2008).
- 83 McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
- 84 Becht, E. *et al.* Evaluation of UMAP as an alternative to t-SNE for single-cell data. *BioRxiv*, 298430 (2018).
- 85 Wang, Z. & Oates, T. in *Twenty-Fourth International Joint Conference on Artificial Intelligence*
- 86 Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemometrics and intelligent laboratory systems* **2**, 37-52 (1987).
- 87 Pal, S. K. & Mitra, S. Multilayer perceptron, fuzzy sets, classification. (1992).
- 88 Noble, W. S. What is a support vector machine? *Nature biotechnology* **24**, 1565-1567 (2006).
- 89 Silva, D. F., De Souza, V. M. A. & Batista, G. E. in *13th International Conference on Data Mining*. 687-696 (IEEE).
- 90 He, K., Zhang, X., Ren, S. & Sun, J. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770-778.
- 91 Shanmugam, D., Blalock, D., Gong, J. G. & Guttag, J. Multiple Instance Learning for ECG Risk Stratification. *arXiv preprint arXiv:1812.00475* (2018).
- 92 Shoeb, A. H. Application of machine learning to epileptic seizure onset detection and treatment. (2009).
- 93 Utz, V. M. *et al.* Presentation of TRPM1-associated congenital stationary night blindness in children. *JAMA ophthalmology* **136**, 389-398 (2018).
- 94 Wang, C. *et al.* Deconvolution of subcellular protrusion heterogeneity and the underlying actin regulator

