Deep Learning and NLP-Based Cardiac Risk Prediction Using ECG and EHR Data

by

Tahmid Iqbal
21201701
Imranul Hasan Emon
20301142
Mehedi Hasan
21201270
Ahtesham Ibne Mustafa
21201342

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
October 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Tahmid Iqbal
21201701

Imranul Hasan Emon
20301142

Mehedi Hasan
21201270

Ahtesham Ibne Mustafa
21201342

Approval

The thesis/project titled "Deep Learning and NLP-Based Cardiac Risk Prediction Using ECG and EHR Data" submitted by

1. Tahmid Iqbal(21201701)
2. Imranul Hasan Emon(20301142)
3. Mehedi Hasan(21201270)
4. Ahtesham Ibne Mustafa(21201342)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on October 19, 2024.

Examining Committee:

Supervisor:
(Member)

Name of Supervisor
Designation
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

With the tremendous development of deep learning (DL) and natural language processing (NLP), great potential is offered in predicting cardiac risk. We have proposed deep learning and NLP-based cardiac risk prediction using electrocardiogram (ECG) and electronic health record (EHR) data. The study aims to apply such technologies in the prediction of cardiac risks through the analysis of electrocardiograms and electronic health record data. To this end, we would like to to propose a DL-based multimodal architecture for processing the 12-lead ECG signals in tandem with NLP techniques on comments and cardiologist reports. We go further to expand this dual approach to realize our objective of improving accuracy and reliability in cardiac risk assessment. Regarding this, our study leverages data from the MIMIC-IV-EGG module, which includes about 800,000 diagnostic ECGs from nearly 160,000 unique patients. These ECGs are sampled at 500 Hz and span 10 seconds each; they are linked to the huge MIMIC-IV Clinical Database, integrating extensive patient data on demographics, diagnoses, medications, and lab results. Our approach has been to feed raw ECG signals into a deep-learning model, predict cardiac events, and generate cardiological reports. These comments, therefore, undergo further analysis with the help of LLMs so that potential diseases can be identified and refined cardiac risk predictions can be generated. Such a novel approach helps in building a vigorous predictive model that accurately identifies cardiac risks and provides detailed cardiological reports. Thus, our model enhances efficiency in the protection of cardiac health.

Keywords: Deep-Learning, Large Language Model (LLM), Natural Language Processing, ECG, EHR.

Acknowledgement

To begin with, we would like to give our heartiest gratitude first and foremost to the Great Allah for showing us the way to complete our thesis, Phase-1 without significant interruptions.

We are also deeply and profoundly grateful to our supervisor, Dr. Md. Ashraful Alam, for his valuable support and encouragement during our work. His help was very much materialized any time we were in problems.

Not the least, special thanks go to our parents for their support whenever we needed them. Their consistent support and prayer have been a strength to us right through to this moment when we are about to graduate.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

Globally, Cardiovascular disorders are recognized as the primary cause of mortality, accounting for the loss of millions of lives, annually. It is now necessary to identify heart diseases as early as possible and forecast cardiac risks with high accuracy so that better patient outcomes can be assured and adverse cardiovascular events avoided. The technologies of AI, especially deep learning and natural language processing, have recently gained considerable attention. Interest is increasing in using such deep learning and natural language processing methods for further improvement in cardiac risk prediction models. Parsing through dense medical information and connections that may otherwise remain invisible to more classic approaches can now afford better risk prediction.

These two major sources of information involve ECG data and electronic health records for assessment. ECGs yield detailed information concerning the electrical activities of the human heart; thus allowing the detection of arrhythmias, ischemia, as well as other cardiovascular conditions. While structured data in EHRs is represented by demographics, diagnosis, medication, and lab results, unstructured data involves clinician notes and cardiology reports. While ECGs represent significant real-time information on the electrical functioning of the heart, EHRs provide a broad overview of a patient's medical history and hence are very much crucial for long-term cardiac risk assessment.

Despite the amount of information in both ECGs and EHRs, most of the existing models for predicting cardiac risk do face a number of limitations. Most of the models are based on either ECG or EHR data alone; there is no integration of both into one holistic view of a patient's cardiovascular health. Besides, the traditional models lack the processing and interpretation of unstructured clinical data in physician notes, which bear important insights. Another possible limitation of the current approaches is small sample sizes, which may limit generalization and robustness in various patient populations. In this regard, the challenge is tackled by proposing a multimodal approach through a combination of DL and NLP techniques to perform an analysis of both ECG signals and EHR data for cardiac risk prediction.

1.2 Approach

Our approach leverages the MIMIC-IV-ECG module, which adds approximately 800,000 diagnostic ECGs to the extensive EHR data available from the MIMIC-IV Clinical Database, utilizing deep learning architectures for processing 12-lead ECG signals and NLP models for analyzing clinical notes and cardiologist reports. It integrates two data modalities into one, enabling researchers to make more precise cardiac risk assessments by permitting the simultaneous analysis of the electrical activities of the human heart, along with a patient's overall medical profile.

In addition, our model tries to address scalability and efficiency challenges in a real-time clinical setup through optimized computational needs and enhanced interpretability of the model. The research study's objective would be to provide an integrated, reliable, and scalable model for cardiac risk prediction that can be deployed in clinical environments to help with early diagnosis and proactive management of cardiovascular conditions.

In the subsequent sections, we discuss related work, the research problem, methodology, results, and conclusion of this study, focusing on how our proposed multimodal approach can improve the accuracy and clinical applicability of cardiac risk prediction using DL and NLP techniques.

1.3 Motivation

The motivation for this work is the disturbing global burden of cardiovascular affliction, still the major reason for death and taking the lives of millions every year. The traditional risk prediction models mostly suffer from a lack in their accuracy and applicability due in large part to their reliance on either ECG data or EHRs exclusively. This disparity not only limits their effectiveness in clinical settings but also delays intervention, which is very critical for improving patient outcomes. We aim to harness the power of DL and NLP to develop a more complete and integrated model that encompasses both structured and unstructured data. We therefore try to decode the underlying connection between ECG signals and detailed patient medical history captured through EHRs that might pass on undetected through traditional means. The current study intends to make a good contribution to cardiology by enhancing the accuracy in risk prediction leading to early diagnosis and active management of cardiovascular diseases, reducing the healthcare burden across the world.

Chapter 2

Problem Statements and Objectives

2.1 Problem Statements

A lot of research has been done so far on cardiac risk prediction with ECG and structured as well as unstructured EHR datasets. The authors tried to get a higher accuracy rate by applying various models like Self Supervised Learning(SSL) models, Long Short Term Memory(LSTM), Bidirectional Long Short Term Memory(BiLSTM), Convolutional Neural Networks (CNNs) etc. However, maximum existing cardiac risk prediction models suffer limitations like single datasets (such as either ECG or EHR data) or small datasets. Some prediction models might fit well on these datasets and provide impressive results but these small datasets could have an impact on the model's adaptability. Therefore, these current models often struggle to provide a complete clinical picture of a patient's cardiovascular health. Moreover, some existing models fall short when it comes to understanding and analysing unstructured clinical data which have insightful information such as patients' written notes and documents from doctors.

In this paper, we intend to fill these research gaps. Our study contributes to these aspects by introducing an effective and comprehensive deep learning and natural language processing based model which helps to predict cardiovascular diseases better than previous models. To do so, we integrated ECG and EHR data and applied our model to them. We employed the technique of Natural Language Processing to extract a significant amount of information from cardiologist reports which are available in our huge MIMIC-IV Clinical Database. At the same time, we analyzed the ECG signals using deep learning techniques. This model is developed to account for both the structured and unstructured data to facilitate a comprehensive and reliable development of a cardiac risk prediction model.

2.2 Research Objectives

By integrating deep learning with NLP, this research will be directed at a robust multimodal model to accurately forecast cardiac risk by analyzing 12-lead ECG signals and associated electronic health record data. Precisely, our study's objectives are to:
i. Design and implement a DL architecture so that it can process raw ECG signals in order to predict possible cardiac events.
ii. Use NLP on cardiologists' reports and comments and on the general trends of the EHR data to further develop and improve cardiac risk assessments.
iii. Merge the MIMIC-IV-ECG module with Clinical Database MIMIC-IV into one complete dataset, which would be more robust for training and validation.
iv. Performance and accuracy analysis of the proposed model in identifying cardiac risk and formulating detailed, actionable cardiological reports.
Meeting these objectives will ensure the research comes up with a powerful automated tool that will enhance early detection of cardiovascular diseases to support the better management of the patients for improved outcomes.

Chapter 3

Literature Review

Khurshid et al.[6] (2021) explored strategies to improve the very early detection of AF, one of the most prevalent yet generally asymptomatic cardiac conditions that produce significant health risks, including stroke. Current clinical models, such as CHA2DS2-VASc, have limited predictive power. Dealing with this, a model was proposed that implements deep learning on ECG data by using CNNs scanning minute patterns in the signals emanating from ECG, hinting toward future AF risk. The improvement over their model came in terms of creating a hybrid version: CH-AI, which combined ECG-based predictions with risk-associated variables, hence being way more accurate in predictions than classic models like CHARGE-AF. Further, this study was the first to make use of enormous ECG datasets ranging from three different sources such as the UK Biobank and the Brigham and Women's Hospital databases. Additionally, the authors preprocessed and labeled ECG data into both training and validation sets. Similarly, a type of deep learning model that involves CNNs was used by the authors in this experiment, suitable for both time-series data and waveform data e.g. ECG analysis. It has also been trained to identify patterns, even when there is a normal sinus rhythm on the ECG, which might indicate a high risk for AF. Moreover, the researchers have proposed a hybrid model, CH-AI, that integrates clinical risk factors into the deep learning ECG-based models, ECG-AI, in an attempt to improve prediction accuracy. Because of this method, the accuracy was higher compared to clinical or ECG data models. Moreover, saliency maps enhanced the model's decision-making procedure by highlighting features considered important in ECGs. In this regard, the research study established that the model using deep learning outdid the traditional clinical methods in improving the capability of risk prediction for AF, as well as opening a wide range to possible clinical implementation. However, limitations in generalization across populations and studies that rely on retrospective data are only two of a multitude of issues that will make further research necessary in the years to come.

The authors, Ullah et al.[4] (2021), proposed a process that incorporates the cardiac ECG data for classification in enhancing arrhythmias. In this research work, the authors set out to overcome the disadvantages of previous ECG signal analysis methods, which most often failed either in precision or accuracy. Similarly, their very process involves splitting 1D ECG signals with respect to Q-wave peak timings to produce 2D grayscale images (512x512). With this change, classification can now be done using a 2D Convolutional neural network, abbreviated as CNN. The

progression is from augmentation of data to increase variety in training data to a CNN architecture that includes many convolutional layers with the ReLU activation, several pooling layers for the extraction of features, as well as the layers that are completely linked or connected fully for classification. We have tuned the model with cross-entropy loss, used an Adam optimizer for back-propagation, and included validation procedures to avoid overfitting.

Moreover, the outcomes from the research show that this strategy greatly enhances accuracy in classification over older approaches. arguably the most significant achievement is the introduction of a scalable, effective, and efficient approach for ECG analysis, resulting in the potential to revolutionize applications in medicine.However, the manuscript also recognizes its limitations, highly depending on high-quality ECG data, and calls for future studies to investigate further with more complex architectures, including additional physiological signals to improve generalizability. The study, in conclusion, emphasizes the potential good that deep learning may offer in improving cardiac healthcare interventions.

Galloway et al.[2] (2019) developed a DLM, which they specifically applied to identify hyperkalemia from the analysis of electrocardiogram data obtained from persons with chronic kidney disease. The authors focused on two primary leads because they are effective in identifying signals of height that indicate a level of potassium higher than normal. After that, the model was developed using convolutional neural networks along with recurrent neural networks to efficiently detect both the spatial and temporal patterns present in the ECG data. Thousands of ECGs were collected from a collaboration between the Mayo Clinic and AliveCor; these were matched against serum potassium levels.

The researchers used a five-fold cross-validation approach to improve the efficiency and the precision of the model, applying various signal processing techniques for noise reduction and normalization on the ECG inputs. That means training several variants of DLM with different hyperparameters to enhance the settings. For their evaluation, classic metrics of performance which include AUC, sensitivity, and specificity were considered. The best had an AUC ranging from 0.853 to 0.883, with sensitivities ranging from 88.9% to 91.3%. That would mean that this model may find hyperkalemia from ECG data pretty well. The authors also comment on how the false positives can be reduced, elaborating on better ECG-based exclusion criteria for some conditions like left ventricular hypertrophy. Other studies involve further validation of the model for domestic use in reliability at a clinical setting.

Vaid et al.[7] from their study found that the performance of deep learning models at quantifying biventricular function from electrocardiograms was obtained from the widespread cohort of socioeconomically diverse patients from the Mount Sinai Health System. The authors initiated their study by collecting a large amount of data comprising 700,000-ECGs from over 150,000 unique patients, with the aim of developing models that could predict with high accuracy both left and right ventricular contractile states. They performed a multimodal analysis, where data from ECGs were combined with those from echocardiography, using a sophisticated NLP data analysis pipeline that extracted outcomes from free-text echo reports. Among

the results of these tests, their deep learning model showed better performance not only in classifying the LVEF but also in detecting right ventricular dysfunction, with an AUROC as high as 0.84 for recognizing the impaired state of the right ventricle. Later, the authors produced a regression framework to estimate LVEF using an average error of 5.84%., thus further increasing the clinical value of ECGs as heart failure screening tools. This study demonstrated that their models were truly robust and generalizable, showing that DL techniques could substantially enhance the assessment of cardiac function in patient management within heart failure scenarios.

Liu et al.[1] (2018) investigated the possibility of predicting chronic diseases using deep learning algorithms; their work was based on electronic health records and titled "Deep EHR." Their work focused on improving the accuracy and interpretability of the models; much previous work had lacked transparency, which is a vital aspect that clinicians require. Integration of various types, such as numerical and textual information, was underlined as one of the biggest challenges in which high precision had to be maintained. Such authors tested several DL architectures like CNN and LSTM models on the prediction of diseases like congestive heart failure and stroke. The methodology they employed integrated structured data with unstructured medical documentation through the implementation of an innovative negation tagging technique to address the negation present in medical texts, and utilized regular expressions to derive quantitative variables from the data.

Their process included both structured data integration, such as numerical values, and unstructured ones like medical notes for model training. They used the CNN models for extracting global features from the data; they tested two different interpretability methods- a gradient-based approach and a log-odds approach-showing greater interpretability for the latter. This study reported strong predictive performance, with overall high AUC scores, which indicated that the CNN model was able to pick out key clinical features such as medical conditions and lifestyle factors. It is because the model performance was so good that issues of precision and computational intensity arose. Further improvements in precision, negation tagging enhancement, and enlargement of the model to be able to predict a greater range of diseases are ways this study can be taken further. This investigation follows and further develops other current efforts in the domain of clinical prediction by using EHRs, where a key balancing challenge between accuracy and interpretability remains open.

In their study, J. Weston Hughes and his colleagues developed a hybrid deep learning model (CH-AI) using Convolutional Neural Networks integrated with traditional clinical risk factors to predict long-term cardiovascular outcomes such as conditions associated with the cardiovascular system including carditis, aorta aneurysms, peripheral vascular disease, embolic diseases, venous thrombotic episodes, congenital heart defects, dysfunction of heart valves, symptoms of heart failure, hypertensive changes in the heart, rheumatic diseases of the heart, myocardial diseases, abnormal rhythm in heart contractions, coronary vascular diseases such as angina and myocardial infarction, and peripheral vascular disorders. When ECG signal analysis and clinical data were combined, the model significantly outperformed previous atrial fibrillation (AF) prediction methods. A notable feature is using saliency maps

for interpretability, enhancing clinical trust. Their future research aims to explore real-time applications, wearable device integration, and validation across diverse populations for broader generalizability and continuous monitoring.

Yaqoob Ansari and colleagues in their paper present the overview of DL models for ECG-based detection of arrhythmia and classification developed during the period from 2017 up to 2023. It highlights models such as CNNs, RNNs, and Transformers for their advanced capabilities. The survey offers guidelines for new researchers, focusing on trends, optimizing model architectures, and handling large-scale ECG data. Future improvements include refining DL models for better clinical applications and addressing the challenges of large-scale data processing. The paper serves as a roadmap for those developing DL models in cardiology.

García-Ordás et al.'s research "Heart Disease Risk Prediction Using Deep Learning Techniques with Feature Augmentation" proposes a new multitask neural network model for improved forecasting of heart disease risks. The researchers have used a Sparse Autoencoder (SAE) to augment features and a classifier, including CNNs or Multilayer Perceptrons (MLPs), to perform classification. This approach enhances predictive accuracy by extracting deeper, more meaningful features. Classic methods like decision trees and random forests are also discussed for comparative purposes, with the multitask model demonstrating superior accuracy. The authors suggest refining feature augmentation techniques and expanding model generalization for future research, highlighting its potential for real-world clinical applications.

The study by Gustafsson et al. presents a significant advancement in the diagnosis of myocardial infarction (MI) in emergency department settings by developing and validating a deep learning model that utilizes electrocardiograms (ECGs). Their model, which classifies patients into NSTEMI, STEMI, and control categories, was presented based on a huge research database of 492226 ECGs from 214250 patients within the Stockholm region, drawn between the years 2007 and 2016. It should be noted that the primary algorithm incorporated also worked on patients' predictors e.g. age and gender with the ECG data. A combination of five models achieved good accuracy, with C-statistics reaching 0.832 for NSTEMI and 0.991 for STEMI, emphasizing the potential of the model as a clinical decision support system. The authors achieve what several previous studies conducted in more controlled environments have not dared to do by stressing the importance of real-world datasets. Future work will focus on improving model generalizability across multiple countries and embedding model usage as a standard practice in the clinics, which, in turn, presents a significant improvement on its own towards early diagnosis and treatment of MI and bettering patient care in emergencies.

The work done by Houssein, Mohamed, and Ali leverages advanced NLP techniques to study the automatic detection of heart disease risk-variables from clinical writings. It also addresses the limitations of previous research that relied on labor-intensive hybrid systems while struggling to extract all relevant risk attributes. In the traditionally complex architectures and instead of the often utilized BERT, CHARACTER-BERT Embedding a sub-character model is used in the proposed framework. This methodology enhances the model's ability to associate relevant

tags and attributes of heart disease risk factors. The i2b2 challenge dataset serves as the primary dataset to demonstrate the effectiveness of their approach, which achieves a remarkable F1 score of 93.66%, surpassing previous systems used in the challenge. Future work is clearly defined in that, the authors will enhance these detection techniques and also investigate other areas where their methods could be used in the clinics, pointing out that the risk factors and the amount of care given to patients in the health care systems could be improved.

This study, entitled "Natural Language Processing to Improve Prediction of Incident Atrial Fibrillation Using Electronic Health Records"[5] , investigates the application of NLP in improving the prediction of AF risk in older adults through narrative text analyses from EHRs. They developed two predictive models: one based solely on codified EHR data and another that combined codified data with information extracted using NLP. Results across more than 86,000 patients showed that the NLP-enhanced model demonstrated superior predictive performance compared with both the codified-only model and traditional CHARGE-AF score. Both models exhibited good calibration in internal validation but overestimated risk in external validation cohorts. Results therefore indicate that NLP would better identify high-risk patients, which would allow for targeted AF screening and prevention strategies. However, this needs further validation in various settings.

The study entitled "Cohort design and natural language processing to reduce bias in electronic health records research" addressed bias in EHR-based studies by establishing the Community Care Cohort Project, where over 520,000 patients receive primary care. To reduce the amount of missing data, they employed NLP to recover vital signs from clinical notes that were unstructured, thereby reducing missingness by 31%. Comparing C3PO to the convenience samples showed better calibration of the risk models for heart disease and atrial fibrillation, highlighting reduced bias. For the first time, the authors introduced the JEDI pipeline for processing diverse EHR data, providing a scalable tool for future research. This work improves both the accuracy and generalizability of healthcare research and may have an impact on public health.

The study "LLMs-based Few-Shot Disease Predictions using EHR: A Novel Approach Combining Predictive Agent Reasoning and Critical Agent Instruction" has looked into the application of large language models to predict diseases with electronic health records, considering that their traditional methods of supervised learning lacked large labeled data. The research team now proposes a new framework called EHR-CoAgent, which involves two collaborative LLM agents: a Predictor Agent that makes the predictions of the disease along with the reasoning behind those predictions, and a Critic Agent that assesses any predictions which are not correct to provide feedback so that further improvements can be made. Structured EHR data are transformed into natural language narratives to explicitly allow for the capability of LLMs to apply their effective reasoning. It also examines the performances of Zero-Shot and Few-Shot LLMs on datasets such as MIMIC-III and CRADLE using various prompting strategies. Results are promising and show that EHR-CoAgent significantly improves diagnostic performance compared to traditional machine learning models, even in few-shot settings. In general, the

results highlight the role of LLMs as useful assistants in clinical decision-making for healthcare, and in particular, in the accurate prediction of diseases with minimal training data.

The study[8], "Large Language Model-informed ECG Dual Attention Network for Heart Failure Risk Prediction", a novel ECG-based dual-attention network that predicts heart failure risk in light of the face of rising global HF mortality, capturing intricate cross-lead interactions and local dynamics of 12-lead ECGs for enhanced interpretability and better predictive accuracy. To mitigate limitations in the available data and get a better learning of features, the network is pre-trained using the large language model with ECG reports by aligning ECG features with clinical knowledge. Further, this is fine-tuned on two UK Biobank cohorts of hypertensive patients and those with a history of myocardial infarction. Among these, the proposed approach achieved the best performance, with high C-index scores of 0.6349 for hypertension and 0.5805 for myocardial infarction, outperforming traditional methods. This reflects great potential to further improve early detection and risk stratification of HF. Moreover, this model is further improved in interpretability with a dual attention mechanism, increasing its clinical value.

The study[11], "ECG Semantic Integrator (ESI): A Foundation ECG Model Pre Trained with LLM-Enhanced Cardiological Text", proposes the ECG Semantic Integrator, namely ESI, a deep learning-based multimodal framework, to further advance the analysis of ECGs. ESI enhances the semantic comprehension of 12-lead ECG signals through two means:. CQA relies on LLMs and the knowledge of medicine which can generate enriched textual-descriptions of ECG, including demographic and waveform information. The ECG Semantic Integrator then couples the ECG signals with these textual descriptions via contrastive and captioning losses, pretraining the ECG encoders and giving rise to more robust signal representations. It has been tested on large datasets and shows significant improvement over the existing methods of supervised and self-supervised fashion on tasks such as arrhythmia detection and patient identification, thus showing the potential of multimodal learning combined in enhancing cardiac diagnostics.

Insufficient diagnostic coding in claims data is the foundation of our current methods for identifying hospitalisations for worsening heart failure (WHF). In their work, Ambrosy et al.[3] (2021) discuss the limitations of our current approach and demonstrate how underreporting leads to a mischaracterization of the actual burden of WHF. The study intends to increase the precision of WHF detection by using the NLP algorithm for both structured and unstructured electronic health records (EHRs). The authors employed natural language processing (NLP) methods to analyze pertinent free text notes found in the electronic health record (EHR) within the 72 hours preceding an admission. A regular expression technique was used to separate all the notes to detect line endings and specified section headings. Then I2E software(version 6.2.0) was used to extract relevant information by applying custom ontologies to capture accurate clinical data. Two doctors reviewed and validated the NLP algorithms manually in order to verify them to a criterion standard. The model was tested against manual record reviews and it shows excellent sensitivity (98%) and specificity (95%). Compared to our current existing approaches that

solely rely on principal discharge diagnoses, the detection of WHF hospitalisations is doubled when the NLP algorithm is applied. During the study period, hospitalisations for WHF increased significantly, especially among those who have HFpEF. These results indicate that WHF is a far more common issue than we previously thought.

Combining rule-based natural language processing (NLP) with both unstructured and structured data offers a more thorough and precise approach to WHF identification. In contrast to earlier studies which mostly relied on claims data, this method makes use of real-time EHR data to increase sensitivity without compromising specificity. This study not only finds more WHF cases but also reveals clinical subgroups and temporal trends, particularly in underreported populations like those who have HFpEF. The study has some limitations. Because of inconsistent documentation, it does not include long-term therapies such as dialysis and IV vasoactive medicines. In order to distinguish acute WHF from issues that develop later in hospitalisation, the study also concentrated on data from the first 72 hours of admission. Another disadvantage is the potential for hospitalisations that began at non-affiliated facilities may go unreported. Future studies should apply NLP to ambulatory care settings in order to reduce reliance on hospital-based data alone and better capture the patient's journey throughout several healthcare environments.

In a research, Nagamine et al. (2022) divided HF patients into 23 different disease states using unsupervised clustering by K-means based on complaints taken from unstructured EHRs. Patient snapshots were vectorised using TF-IDF, which aids in identifying and classifying clusters and comparable disease states. Using a big dataset, this study examines the complex nature of HF. The authors show diverse HF illness states by clustering patients into temporal snapshots of their data. Instead of structured data, the analysis of unstructured clinical notes through natural language processing (NLP) improves the dataset with clinical complaints, symptoms and patient outcomes. This provides greater insights and information into the course of heart failure. The focus of this study is to enhance our understanding and knowledge of heart failure as a disease by identifying particular disease states and evolution paths which are often ignored in our traditional classification methods.

Pachiyannan et al.[9] (2024) propose an automated detection methodology to improve the accuracy of congenital heart disease (CHD) identification. The authors have proposed the ML-CHDPM machine learning model for the early diagnosis of congenital heart disease which uses ECG signals as input. In their model, they used attention Mechanisms (AMs), Convolutional Neural Networks (CNN), and Bidirectional Long Short-Term Memory (BiLSTM) networks. BiLSTM was used to track the temporal dependencies in time-series data while CNNs serve as feature extractors for the ECG signals. Attention mechanisms enhances predictions further paying emphasis on the important aspects of time series. This approach was applied on the database obtained from the ECGs of the pregnant women. The model proposed in this paper attained an accuracy rate of 96.51% which is higher than just CNN, Random Forest (RF) and Naïve Bayes (NB) methods. Furthermore, the accuracy, recall, and specificity measures performed better than other methods which are already in use with an average precision of 89.14% and recall of 99.19%. The use of complex

neural network architectures like CNN, BiLSTM, and Attention Mechanisms to a particular dataset consisting of ECGs of pregnant women makes their article notable. This method not only fills a research gap on cardiovascular problems during pregnancy but also improves the diagnosis accuracy of CHD. The primary limitation of this study is that it focuses only on ECG signals. Other data such as genetic or lifestyle information could enhance the predictions even further. In future work, researchers can work on a more diverse and larger dataset. Additionally, this model would be more practical for real-time healthcare situations by improving computing efficiency and shortening training time.

In a study by Sattar et al.[10] (2024), a digital dataset made from ECG images collected from multiple medical facilities is used where the authors used some deep learning models like Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks and Self-Supervised Learning (SSL) models to their collected data. CNN is the primary model in their study because of its capacity to recognise spatial and temporal patterns in ECG data. An open-source software is employed that transforms the ECG images into time-series signals in order to fit ECG images into their deep learning model. The CNN is intended to analyze features of the ECG signals and classify them into the four groups. Those four groups are normal, myocardial infarction (MI), abnormal heartbeat (AHB) and prior MI (PMI). Each of the groups contains unique features. The classification accuracy achieved by the presented CNN model nearly 92% . This high accuracy exceeded all the other models tested including LSTM and SSL. High accuracy reported by the CNN model indicates the possibility of using it for the detection of arrhythmias from ECG signals. This model's effectiveness can also be attributed to a fast inference time of 0.1 seconds per prediction and making it possible for real time clinical monitoring. This study demonstrates the ability of CNN to analyze complicated ECG signals and produce accurate results. But there are some limitations also. The authors uses a very small size of dataset (928 samples) in its execution. This small dataset may also affect the flexibility of the model although the authors employed SMOTE to alleviate this by balancing the class distributions also. Future studies to address this issue may concentrate on either obtaining more ECG data or using data augmentation methods. Also, a diverse dataset should be the focus to enhance the scope of the model in subsequent works.

A study by Daydulo et al. (2023) reveals that doctors can minimize mistakes and the time taken to diagnose if they employ automated deep-learning algorithms. These algorithms are able to quickly detect and categorize any cardiac arrhythmia. This method uses time-frequency ECG data to classify three conditions namely congestive heart failure (CHF), cardiac arrhythmia (ARR) and normal sinus rhythm (NSR). The main goal of this study was to create an automated and deep-learning-based ECG segmentation system that can precisely classify ECG data into these three conditions. Their approach comprises of preprocessing the information, converting ECG signals into vision time-frequency representations and classifying the information by the use of deep learning models trained beforehand. Their database was collected from PhysioNet particularly the MIT-BIH database and the BIDMC database. 1D ECG time-series data was enhanced into 2D images using the Morse wavelet. Three categories were created by refining the pre-trained AlexNet and ResNet50 models to

classify the ECG data. The accuracy and resilience scores of ResNet50 were higher than those of AlexNet. Hyperparameters including a batch size of 30 and a learning rate of 0.0001 were used in the training process along with batch normalisation and the Adam optimiser. The proposed deep learning model in this study performs exceptionally well in categorisation. Because the model's overall accuracy was 99.2%. with an average F1-score of 99.%, sensitivity of 99.2%. and specificity of 99.6%. The authors achieve better results than previous models like AlexNet. This work shows that using deep learning models such as ResNet50 in conjunction with Morse wavelet-based time-frequency representations may detect cardiac arrhythmias from ECG data with remarkable efficiency and accuracy.

Chapter 4

WorkPlan

Our intention here is to elaborate on the ECG and EHR data available within the MEMIC-IV database, with special attention on the signal and text preprocessing techniques. To begin with, raw ECG signals are subjected to signal preprocessing which incorporates basic steps of noise filtering, physiological signals artificially removing the non-physiological interferences, data shaping into segments of manageable sizes, data scaling, and aligning the signals through re-referencing. On the other hand, the EHR which is usually in the prose format also undergoes text cleaning before being analyzed using text processing techniques involving conversion of uppercase letters to lowercase, segmenting or breaking down the text into the smallest possible words, eliminating auxiliary words and lemmatization or stemming which entails the reduction of a specific word to its base or root. To cope with the lack of data, i.e. introduce imputation, use resolving discrepancies between variable types, and adjust the length of data entries so that all data is the same length, by either cutting excess data or adding additional irrelevant data to the shorter entries. After data processing, ECG and EHR data are fused, and features a built on top of this data to extract informative attributes, for example, for training supervised machine learning, deep learning models, or reinforcement learning. The following steps include selecting a model, training this model, and adjusting hyperparameters for the chosen algorithm. The last stage consists of preparing the resulting model for deployment which enables its use for machine learning model usage to help identify outlying heart signals given the processed EHR and ECG data.

Here is the simple work flow chart given.

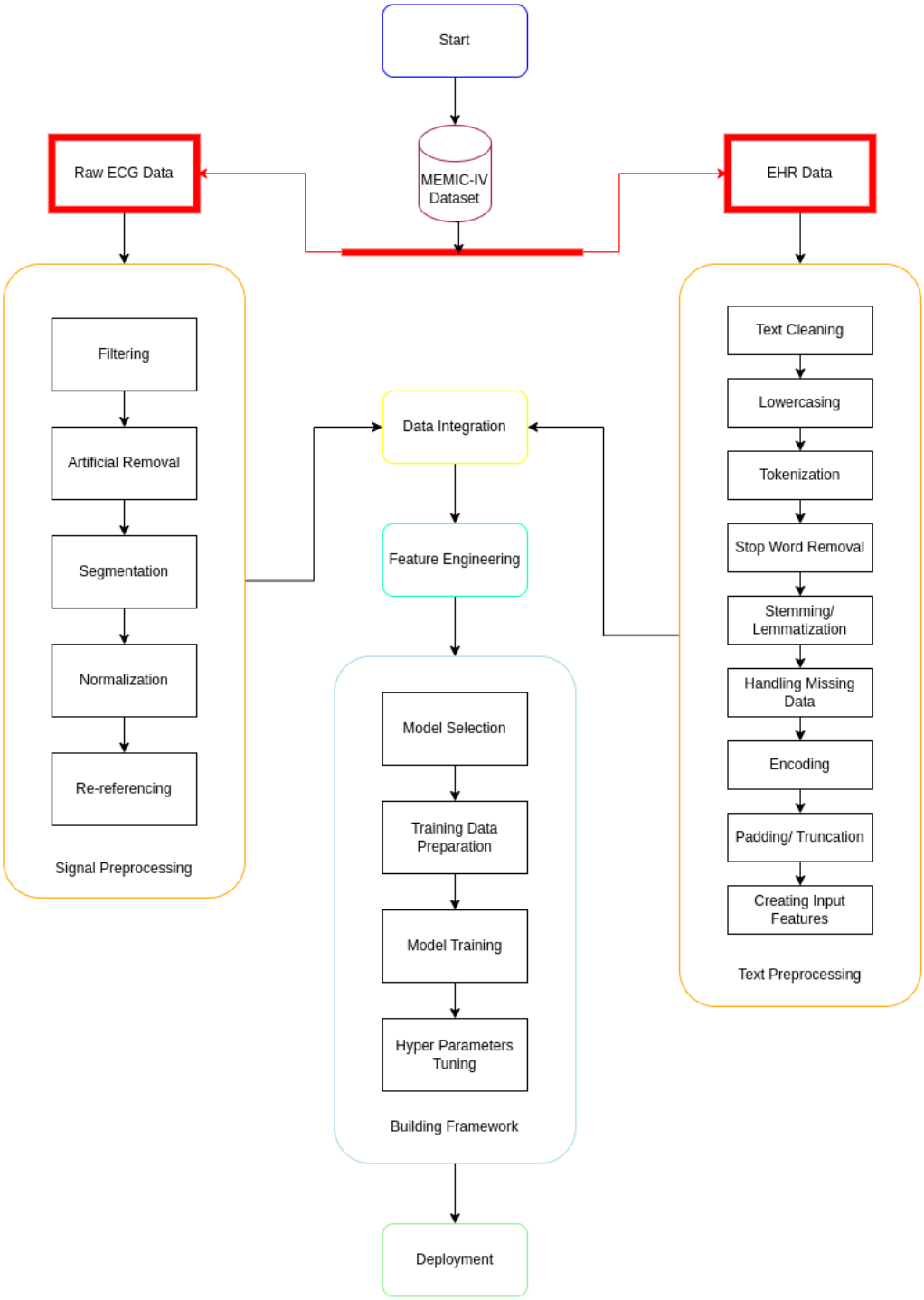

Figure 4.1: Workflow

Chapter 5

Conclusion

Since cardiovascular diseases are one of the leading causes of death in this age, it is crucial to identify cardiac risks early and accurately. Application of deep learning and natural language processing based algorithms in a large and diverse dataset can develop a good cardiac risk prediction model which can be able to predict more accurately and efficiently than previous models. That is why our study seeks to develop a system that can better predict cardiac risks by combining unstructured data from EHRs with ECG signals after applying our dual deep learning-based model to it. By merging these two data sets, we intend to develop a model that can enhance risk assessments' accuracy and comprehensibility. This study shows how Natural Language Processing(NLP) and deep learning(DL) can work together to improve cardiac risk prediction and provide medical professionals with a vital tool.

Bibliography

[1] J. Liu, Z. Zhang, and N. Razavian, "Deep ehr: Chronic disease prediction using medical notes," in *Proceedings of the 3rd Machine Learning for Healthcare Conference*, F. Doshi-Velez *et al.*, Eds., ser. Proceedings of Machine Learning Research, vol. 85, PMLR, 17–18 Aug 2018, pp. 440–464. [Online]. Available: https://proceedings.mlr.press/v85/liu18b.html.

[2] C. D. Galloway *et al.*, "Development and Validation of a Deep-Learning Model to Screen for Hyperkalemia From the Electrocardiogram," *JAMA Cardiology*, vol. 4, no. 5, pp. 428–436, May 2019, ISSN: 2380-6583. DOI: 10.1001/jamacardio.2019.0640. eprint: https://jamanetwork.com/journals/jamacardiology/articlepdf/2729582/jamacardiology_galloway_2019_oi_190012.pdf. [Online]. Available: https://doi.org/10.1001/jamacardio.2019.0640.

[3] A. P. Ambrosy *et al.*, "A Natural Language Processing–Based Approach for Identifying Hospitalizations for Worsening Heart Failure Within an Integrated Health Care Delivery System," *JAMA Network Open*, vol. 4, no. 11, e2135152–e2135152, Nov. 2021, ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2021.35152. eprint: https://jamanetwork.com/journals/jamanetworkopen/articlepdf/2786416/ambrosy_2021_oi_210989_1636735785.14819.pdf. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2021.35152.

[4] A. Ullah, S. u. Rehman, S. Tu, R. M. Mehmood, Fawad, and M. Ehatisham-ul-haq, "A hybrid deep cnn model for abnormal arrhythmia detection based on cardiac ecg signal," *Sensors*, vol. 21, no. 3, 2021, ISSN: 1424-8220. DOI: 10.3390/s21030951. [Online]. Available: https://www.mdpi.com/1424-8220/21/3/951.

[5] J. M. Ashburner *et al.*, "Natural language processing to improve prediction of incident atrial fibrillation using electronic health records," *Journal of the American Heart Association*, vol. 11, no. 15, e026014, 2022. DOI: 10.1161/JAHA.122.026014. eprint: https://www.ahajournals.org/doi/pdf/10.1161/JAHA.122.026014. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/JAHA.122.026014.

[6] S. Khurshid *et al.*, "Ecg-based deep learning and clinical risk factors to predict atrial fibrillation," *Circulation*, vol. 145, no. 2, pp. 122–133, 2022. DOI: 10.1161/CIRCULATIONAHA.121.057480. eprint: https://www.ahajournals.org/doi/pdf/10.1161/CIRCULATIONAHA.121.057480. [Online]. Available: https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.121.057480.

[7] A. Vaid *et al.*, "Using deep-learning algorithms to simultaneously identify right and left ventricular dysfunction from the electrocardiogram," *JACC: Cardiovascular Imaging*, vol. 15, no. 3, pp. 395–410, 2022, ISSN: 1936-878X. DOI: https://doi.org/10.1016/j.jcmg.2021.08.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1936878X21006276.

[8] C. Chen, L. Li, M. Beetz, A. Banerjee, R. Gupta, and V. Grau, *Large language model-informed ecg dual attention network for heart failure risk prediction*, 2024. arXiv: 2403.10581 [q-bio.QM]. [Online]. Available: https://arxiv.org/abs/2403.10581.

[9] P. Pachiyannan, M. Alsulami, D. Alsadie, A. K. J. Saudagar, M. AlKhathami, and R. C. Poonia, "A novel machine learning-based prediction method for early detection and diagnosis of congenital heart disease using ecg signal processing," *Technologies*, vol. 12, no. 1, 2024, ISSN: 2227-7080. [Online]. Available: https://www.mdpi.com/2227-7080/12/1/4.

[10] S. Sattar *et al.*, "Cardiac arrhythmia classification using advanced deep learning techniques on digitized ecg datasets," *Sensors*, vol. 24, no. 8, 2024, ISSN: 1424-8220. [Online]. Available: https://www.mdpi.com/1424-8220/24/8/2484.

[11] H. Yu, P. Guo, and A. Sano, *Ecg semantic integrator (esi): A foundation ecg model pretrained with llm-enhanced cardiological text*, 2024. arXiv: 2405.19366 [eess.SP]. [Online]. Available: https://arxiv.org/abs/2405.19366.

Bibliography

October 19, 2024

1. Khurshid, S., Friedman, S., Reeder, C., Di Achille, P., Diamant, N., Singh, P., Harrington, L. X., Wang, X., Al-Alusi, M. A., Sarma, G., Foulkes, A. S., Ellinor, P. T., Anderson, C. D., Ho, J. E., Philippakis, A. A., Batra, P., & Lubitz, S. A. (2021). ECG-based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation, 145*.

2. Ullah, A., Rehman, S. u., Tu, S., Mehmood, R. M., Fawad, & Ehatisham-ul-haq, M. (2021). A hybrid deep CNN model for abnormal arrhythmia detection based on cardiac ECG signal. *Sensors, 21*, 951.

3. Galloway, F., Valys, A., Shreibati, J. B., Petterson, A., et al. (2019). Using ECG-based DLM for Hyperkalemia Detection. *JAMA Cardiology*.

4. Vaid, A., Johnson, K. W., Badgeley, M. A., Somani, S. S., Bicak, M., Landi, I., Russak, A., Zhao, S., Levin, M. A., Freeman, R. S., Charney, A. W., Kukar, A., Kim, B., Danilov, T., Lerakis, S., Argulian, E., Narula, J., Nadkarni, G. N., & Glicksberg, B. S. (2022). Using deep-learning algorithms to simultaneously identify right and left ventricular dysfunction from the electrocardiogram. *JACC Cardiovascular Imaging, 15*, 395–410.

5. Liu, J., Zhang, Z., & Razavian, N. (2018). Deep EHR: Chronic disease prediction using medical notes. *Proceedings of the 3rd Machine Learning for Healthcare Conference, Proceedings of Machine Learning Research, 85*, 440 464.

6. Hughes, J. W., Tooley, J., Torres Soto, J., et al. (2023). A deep learning-based electrocardiogram risk score for long term cardiovascular death and disease. *npj Digit. Med. 6*, 169.

7. Ansari, Y., Mourad, O., Qaraqe, K., & Serpedin, E. (2023). Deep learning for ECG Arrhythmia detection and classification: an overview of progress for period 20172023. *Front Physiol, 14*, 1246746.

8. García-Ordás, M. T., Bayón-Gutiérrez, M., Benavides, C., et al. (2023).

Heart disease risk prediction using deep learning techniques with feature augmentation. *Multimed Tools Appl, 82*, 31759–31773.

9. Gustafsson, S., Gedon, D., Lampa, E., et al. (2022). Development and validation of deep learning ECG-based prediction of myocardial infarction in emergency department patients. *Sci Rep, 12*, 19615.

10. Houssein, E. H., Mohamed, R. E., & Ali, A. A. (2023). Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques. *Sci Rep, 13*, 7173.

11. Ashburner, J. M., Chang, Y., Wang, X., Khurshid, S., Anderson, C. D., Dahal, K., Weisenfeld, D., Cai, T., Liao, K. P., Wagholikar, K. B., Murphy, S. N., Atlas, S. J., Lubitz, S. A., & Singer, D. E. (2022). Natural language processing to improve prediction of incident atrial fibrillation using electronic health records. *Journal of the American Heart Association, 11*.

12. Khurshid, S., Reeder, C., Harrington, L. X., Singh, P., Sarma, G., Friedman, S. F., Di Achille, P., Diamant, N., Cunningham, J. W., Turner, A. C., Lau, E. S., Haimovich, J. S., Al-Alusi, M. A., Wang, X., Klarqvist, M. D. R., Ashburner, J. M., Diedrich, C., Ghadessi, M., Mielke, J., & Lubitz, S. A. (2022). Cohort design and natural language processing to reduce bias in electronic health records research. *Npj Digital Medicine, 5*.

13. Cui, H., Shen, Z., Zhang, J., Shao, H., Qin, L., Ho, J. C., & Yang, C. (2024). LLMs-based Few-Shot Disease Predictions using EHR: A Novel Approach Combining Predictive Agent Reasoning and Critical Agent Instruction. *arXiv*.

14. Chen, C., Li, L., Beetz, M., Banerjee, A., Gupta, R., & Grau, V. (2024). Large Language Model-informed ECG Dual Attention Network for Heart Failure Risk Prediction. *arXiv*.

15. Yu, H., Guo, P., & Sano, A. (2024). ECG Semantic Integrator (ESI): A Foundation ECG Model PreTrained with LLM-Enhanced Cardiological Text. *arXiv*.

16. Ambrosy, A. P., Parikh, R. V., Sung, S. H., Narayanan, A., Masson, R., Lam, P., Kheder, K., Iwahashi, A., Hardwick, A. B., Fitzpatrick, J. K., Avula, H. R., Selby, V. N., Shen, X., Sanghera, N., Cristino, J., & Go, A. S. (2021). A Natural Language Processin-Based approach for identifying hospitalizations for worsening heart failure within an integrated health care delivery system. *JAMA Network Open, 4*, e2135152.

17. Nagamine, T., Gillette, B., Kahoun, J., Burghaus, R., Lippert, J., & Saxena, M. (2022). Data-driven identification of heart failure disease states and progression pathways using electronic health records. *Scientific Reports, 12*.

18. Pachiyannan, P., Alsulami, M., Alsadie, D., Saudagar, A. K. J., AlKhathami, M., & Poonia, R. C. (2024). A novel Machine Learning-Based prediction method for early detection and diagnosis of congenital heart disease using ECG signal processing. *Technologies, 12*, 4.

19. Sattar, S., Mumtaz, R., Qadir, M., Mumtaz, S., Khan, M. A., De Waele, T., De Poorter, E., Moerman, I., & Shahid, A. (2024). Cardiac arrhythmia classification using advanced deep learning techniques on digitized ECG datasets. *Sensors, 24*, 2484.

20. Daydulo, Y. D., Thamineni, B. L., & Dawud, A. A. (2023). Cardiac arrhythmia detection using deep learning approach and time frequency representation of ECG signals. *BMC Medical Informatics and Decision Making, 23*.