



Heart Disease Prediction Using Late Fusion Over Convolutional Neural Networks

Amel Ksibi¹, Leila Jamel¹, Mohammed Zakariah^{2*}, Manel Ayadi¹, Oumaima Saidani¹, Tagrid Alshalali¹

¹ Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia

² Department of Computer Science and Engineering, College of Applied Studies, King Saud University, Riyadh 11495, Saudi Arabia

Corresponding Author Email: mzakariah@ksu.edu.sa

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.430408>

ABSTRACT

Received: 5 September 2025
Revised: 12 November 2025
Accepted: 12 December 2025
Available online: 31 August 2026

Keywords:

heart disease prediction, deep learning, PhysioNet Challenge 2020 dataset, Late Fusion Over convolutional neural network

Cardiovascular diseases (CVDs) continue to be the leading cause of death worldwide, which calls for the need for innovative, precise diagnostic systems. This paper presents a novel framework for predicting heart disease, utilizing a late-fusion strategy with convolutional neural networks (CNNs). The presented method uses the 2020 PhysioNet/Computing in Cardiology (CinC) Challenge dataset, comprising 43101 12-lead Electrocardiogram (ECG) records of 27 cardiac conditions. By combining deep learning (DL) with late fusion techniques, a model can capture complex temporal and spatial patterns in multiple ECG leads, enabling a network to discriminate between subtle changes in cardiac disorders. Our approach involves training multiple CNN branches on distinct ECG characteristics and fusing their decisions at a later stage to produce the final decision. This approach enhances learning capability by retaining the diversity of individual characteristics' respective representations. The model performed very well in training, achieving an accuracy of 99.86%, a precision of 0.9866, a recall of 0.9857, and an area under the curve (AUC) of 0.9986. However, for validation performance, an accuracy of 95.62% was achieved, indicating a potential for overfitting and further improvements in generalization. Such results support the potential of late-fusion CNNs for predictive cardiology. Future efforts will focus on enhancing the robustness of the models by incorporating additional clinical variables and evaluating domain adaptation methods to improve real-world utility across diverse patient cohorts.

1. INTRODUCTION

Chronic heart failure (CHF) has evolved as not only a prevalent but also an increasing health problem as the heart continuously fails to perform an adequate task of delivering anticipated perfusion of blood to tissue and organ systems (physiologic demands) [1, 2]. The occurrence of CHF is rapidly rising to epidemic levels at 3% per year. CHF has a prevalence in developed nations, affecting about 2–3% of the general population and having a particularly high incidence of 12% in people over the age of 65 [3, 4]. The consequences of CHF are not limited to the field of health, as nowadays, the diagnosis and treatment of this disease account for roughly 3% of the annual costs of health care [5]. This number is an indication of the large economic burden that CHF imposes [6].

Based on these significant health questions, our research is interested in cardiac disease prediction with a special focus on late fusion techniques instead of convolutional neural network (CNN) [7]. Utilizing the insights derived from machine learning (ML), specifically CNN, our research aims to increase the accuracy and efficiency of prediction models for cardiac diseases by analyzing the heart's sound [8, 9]. By applying the methods of late fusion, this means to join the heterogeneous sources of information in order to identify

possibly interacting patterns and increase the accuracy and credibility of forecasts [10, 11].

In electrocardiogram (ECG) signal processing, diverse forms of noise within the acquired signals often undermine the precision of diagnostic data [12, 13]. ECG equipment, renowned for its high sensitivity, frequently encounters a range of interferences, including contact noise, electromagnetic interference (EMI) noise, baseline drift, power line interference, and motion artifacts. Numerous denoising techniques have been extensively investigated on a global scale in response to the urgent requirement for noise reduction [14, 15].

Moreover, implementing a dual-band filter bank alongside a product filter and a lowpass filter to mitigate ECG data noise reflects a notable scholarly achievement [16]. Another approach involves using a deep neural network (DNN) in conjunction with a wavelet scale-adaptive threshold technique and a denoising autoencoder (DAE), resulting in an improved signal-to-noise ratio and root mean square error [17, 18].

An alternative approach uses multi-scale principal component analysis (MSPCA) for efficient removal of noise from the ECG signals prior to processing [19, 20]. While not very well established in the categorization of ECG, the application of MSPCA has shown potential. To eliminate

noise from ECG signals, a technique combining empirical mode decomposition and genetic algorithm threshold technology is used [21]. However, there is awareness of the fact that within the optimization procedure, the genetic algorithm may encounter local optima, which can be a challenge [22, 23]. The utilization of CNNs has been prominent due to their remarkable efficacy in processing multidimensional inputs, such as ECG time series data (1D) and medical images (2D and 3D), demonstrating substantial promise [24, 25]. Within the domain of ECG signal classification, CNNs have shown notable efficacy, exhibiting enhanced accuracy and the ability to extract relevant features from input heartbeats [26] directly.

In addition to ECG applications, deep learning (DL) approaches, particularly CNNs, exhibit exceptional performance in analyzing radiological images, thereby demonstrating their utility and versatility across other medical domains [27, 28]. The combination of multiple disciplines creates an interesting intersection between cardiac disease prediction, ECG signal denoising, and the deep impact of ML and DL in healthcare diagnostics. The primary objective is to improve the accuracy of forecasts and tackle the significant health and economic issues associated with CHF [29].

Recent advances in DL have shown the potential of end-to-end learning, a revolutionary method, in which ML models are able to directly learn from raw data without having pre-identified characteristics [30, 31]. This methodology has proven to be exceptionally efficacious in a wide variety of domains such as pattern recognition, image manipulation, natural language comprehension, speech and audio processing, and sensor data analysis [32, 33]. In the context of

detection of CHF, applying a hybrid approach with both traditional ML methods and end-to-end DL methods exhibits better performance than either method separately.

Whereas the traditional methods of ML heavily rely on a huge number of features that are defined by experts. In contrast, end-to-end DL approaches use learning approaches based on the raw phonocardiogram (PCG) signal in the time domain, and a temporal-domain representation (e.g., the spectrogram [34, 35]). Utilizing a hybrid technique that has been effectively applied in previous research on human activity recognition using smartphone sensor data proves to be a very promising approach to improve the performance and accuracy in CHF and other healthcare domains. The combination of traditional ML, which provides a myriad of ways to learn features, and end-to-end DL, providing ways to see directly from the raw data, is a promising framework to break new ground in the field of healthcare diagnostics.

In the schematic illustration (Figure 1), the way of detecting heart disease using late fusion over CNN is described. The first dataset, the PhysioNet Challenge 2020 dataset, is extracted from the INCART Database, PTB Database, CPSC Database, and the Georgia 12-lead ECG Challenge (G12EC) Database, as well as an unidentified database. Then the elaborate pre-processing phase begins. These include key procedures such as batch generation, class weight calculation, handling of undefined classes, informative processing, one-hot encoding, and preparation of validation data. The Late Fusion over CNN model is then applied to the processed images, which show symptoms of heart disease. Finally, the results undergo several evaluations in model procedures, enabling one to confidently determine whether heart disease is present or not.

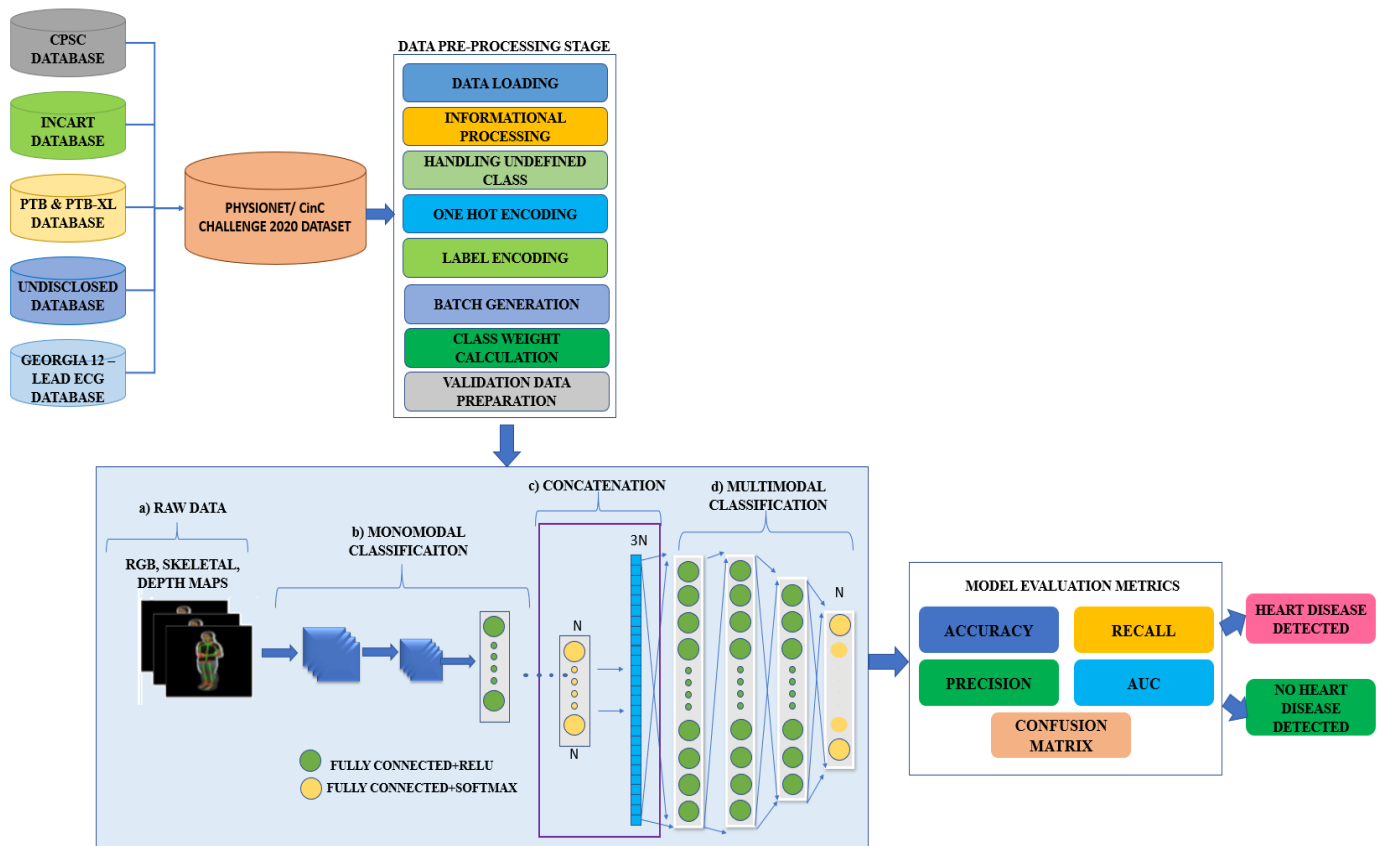


Figure 1. Framework of heart disease prediction using late fusion over a convolutional neural network (CNN)

The Contributions of the research paper are elaborated below:

1. This study adds to the field because it succeeds at integrating the demographic information and the ECG signals so as to predict heart disease. The holistic strategy proposed in this study aims to help manage the various problems encountered in multi-modal data fusion.

2. The study also presents impressive work on optimizing CNNs for maximum performance using the PhysioNet 2020 Challenge dataset. As a result, the introduced process ensures that the model is well-calibrated to the unique characteristics of the provided dataset, which increases its predictive accuracy.

3. The design element of the fusion mechanism involves the introduction of modality-specific streams and the integration of an attention mechanism for fusion. The novelty of this methodology enhances the model's ability to capture relevant characteristics from diverse data sources, thereby improving the overall predictive efficiency of the model.

4. The study focuses on the rigorous optimization of the dataset with proper consideration of the need to tailor the model to the complexity of the PhysioNet 2020 Challenge dataset. Careful tuning of the model improves the efficiency of discovering relevant clinically related patterns related to heart disease.

5. The introduction of the late fusion design is a major improvement, as it allows the model to use data from many ECG leads. It is helpful to perform a well-rounded scan, enabling the model to detect complicated patterns in the context of heart disease across all dimensions of the input data.

The sections below have been presented systematically. Having discussed in detail the existing work done on the subject, Section Two examines that subject matter. Section 3 goes further to explain the various stages that are involved in the data pre-processing and gives a thorough discussion of the datasets used. The study includes the examination of the methods used in Section 4, which gives insight into some of the techniques used for selection. Section 5 gives an extensive account of the experimental setup and the results. After an in-depth discussion of the results in Section 6, Section 7 is dedicated to a coherent conclusion of the paper by briefly summarizing the results and proposing potential directions for further investigation.

2. LITERATURE REVIEW

Over the past decade, DL has surpassed the performance of traditional ML in the interpretation of ECGs with the aid of greater accuracy, eliminating the need for exhaustive feature engineering. The combination of CNNs and Bidirectional Long Short-Term Memory (BiLSTM) improves the classification of ECGs with medical constraints, and improvements were made for the automated diagnosis of cardiovascular diseases [3, 6]. DL has been driving the development in ECG signal analysis for computer-aided diagnosis.

Baumgartner et al. [7] studied privacy-preserving ECG classification for the first time. In their study, they introduce and test six decentralized ML algorithms. The study addresses the relevant issue of assuring the confidentiality of the medical information using the enhanced decentralized learning approach. Building on a 12-lead ECG database, this study employs several methods for model prediction, including

Deep Convolutional Neural Networks (DCNNs), Federated Learning, and Decentralized Learning. Surprisingly, the baseline model has a very good level of accuracy of 91.5%, which serves as a good base for all further research and analysis [8, 9]. However, it is important to note that there are several limitations to the study. More specifically, it only concerns the classification of ECG data. Before expending resources on utilizing this method, one must attempt to resolve the scope and interaction of the technology involved in this approach and test its feasibility in the real world to broaden its rate of application.

In addition, a study on uncertainty quantification approaches for multi-label ECG classification by Barandas et al. [11] highlighted the importance of choosing a dataset as a critical determinant. PhysioNet Challenge 2020 has many 12-lead datasets from many sources (including ECG). Through the application of uncertainty techniques, such as stochastic and ensemble learning, the model accuracy of the research reaches 81.8%. It should be said, however, that the results of this research are rather limited in terms of their applicability since they only apply to the classification of ECGs. It may therefore not be so easy to extend these findings to other medical realms. The suitable measure of uncertainty is still problematic, especially when data sets change. This observation highlights the importance of further research in the important field of automated diagnosis.

In addition, another study carried out by Gjoreski et al. [16] involved the use of machine and end-to-end DL approaches in the detection of CHF on the basis of heart sounds. To obtain relevant information, the authors used seven data sets that included their data sets and data sets that were sent to them from the PhysioNet challenge. The methodology employed for the research included the Random Forest and spectro-temporal ResNet to collect information from a sample of 110 humans who were considered healthy individuals and 51 who were diagnosed with CHF. The accuracy rate of the study was found to be high at 92.9%. However, the scope of this study must be made more expansive, especially in the lacunar range, which only comprises a select set of data currently. It is important to stress the importance of further investigation of the practical issues of implementation and external validation, in order to ensure the widespread adoption of these results.

In addition, Anitha and Rajkumar [21] applied the PhysioNet Challenge dataset to create a logical method for predicting cardiac disease. In this paper, a comparison is made between the extreme gradient boosting (XGBoost) algorithm and CNNs. The obtained balanced accuracy of 89.5% is significantly higher than that of the XGBoost classifiers. As evident from the positive recall values, as well as the area under the curve (AUC), specificity, and test accuracy, this study demonstrates the effectiveness of combining several approaches using the CNN classifier. However, some concerns have been raised, such as issues with the limited scope of the study, which only considered predicting heart disease. External validation and real-world applicability are crucial for a comprehensive assessment.

Similarly, Vasconcellos et al. [26] discussed this problem of inadequate datasets in ECG sets. They particularly focus on intricacies related to 12-lead ECG signals in imbalanced datasets that are burdened with cases of cardiac illnesses. SCNN is a popular method in image classification, and this study aims to utilize this method to classify 12-lead ECG heartbeats using a limited number of training samples. The research primarily focuses on the few-shot learning paradigm

[27, 28]. The SCNN shows an accuracy of up to 86%. Granted, the study points out its limitations by stating that its results are specific to 12-Lead ECG datasets alone. It also highlights the need to generalize these findings to diverse demographics and real-life situations to make them applicable on a broader scale.

Also, the database built by PhysioNet is used by Khan et al. [31] to study ECG categorization. This database consists of forty-eight 2-channel ECG signals recorded from forty-seven subjects. The proposed technique by the author makes use of a one-dimensional convolutional deep residual neural network (ResNet) that learns to extract features directly from the input heartbeats. To solve the class imbalance problem of the class, the synthetic minority oversampling technique (SMOTE) is used on the training dataset in this study, and the average accuracy is improved significantly to 98.63%. It is, however, essential to note the limitations of this study; the vast majority of them are related to the Response community and Response validation and are based on a single ECG database. It is crucial that the need to address real-world implementation issues is recognised to ensure that not only is the expandability, but also the applicability of the answers to the various problem situations in terms of population applicability.

Furthermore, extensive studies by Cheng et al. [35]

highlighted the crucial area of cardiovascular health, with a primary emphasis on the role of ECG signals in diagnosing cardiovascular disorders. Using the PhysioNet Challenge dataset, which contains 8,528 data points for a single lead, this research employs CNN and BiLSTM as the selected methodologies. The model's accuracy rate of 89.3% indicates the effectiveness of classifying the ECG signal using the model. However, the study acknowledges some limitations, including a potential lack of diversity in the datasets, which underscores the need for practical verification to make the model more relevant for various conditions and different populations.

This literature review's findings showed that late fusion methods using CNNs yielded good results in improving heart disease prediction through heart sounds. Challenge datasets are very important in model evaluation. Despite promising results, further work efforts are needed to enhance the applicability and scalability of the model and/or increase its robustness for the development of universally applicable and effective diagnostic tools for broader populations.

Table 1 summarizes the previous references with regard to each of the datasets used, the methods employed, the constraints found, and the results obtained.

Table 1. List of past references, including datasets, approach, limitations, and results

Reference	Dataset	Methodology	Limitations	Results
[7]	- The PhysioNet Challenge 2020 Dataset is used in this paper. - Consists of six 12-ECG-related Datasets.	Deep Convolutional Neural Network (DCNN), Federated Learning, Decentralized Learning	- The study focuses on ECG classification, limiting the generalization of findings to other medical applications. - Technical implementation challenges and real-world feasibility must be carefully considered for widespread adoption.	The baseline model has an accuracy of 91.5%.
[11]	- PhysioNet / Cinc Challenge 2020 is used. - It consists of four 12-ECG datasets	Uncertainty Methods, Stochastic and Ensemble Techniques	- Findings are specific to ECG classification, potentially limiting applicability to other medical domains. Challenges in accurately quantifying uncertainty, especially with dataset shifts, persist.	Uncertainty Accuracy is 81.8%.
[16]	- The Physionet Challenge dataset is used. - It consists of six datasets, containing three thousand one hundred and fifty-three heart-sound recordings.	Machine Learning (ML), deep learning (DL), Random Forest, Spectro-temporal ResNet	- The evaluation focuses on specific datasets, potentially limiting generalization to diverse populations. Real-world implementation challenges and external validation are essential considerations for widespread adoption.	The accuracy of this model is 92.9%.
[21]	- This research utilizes the dataset from the PhysioNet Challenge 2020.	DL, KNN, SVM, Naïve Bayes, XGBoost, convolutional neural network (CNN)	- The study focuses on heart disease prediction, potentially lacking diversity in datasets. External validation and real-world applicability considerations are essential for comprehensive evaluation.	This model achieves an accuracy of 89.5%.
[26]	- ECG Signals are available on the PhysioNet database. - It consists of seventy-five ECG recordings extracted from thirty-two records. - The authors have used the PhysioNet database for ECG signals.	Few-Shot Learning, Siamese Convolutional Neural Network (SCNN), Discrete Wavelet Transform Approach	- The study focuses on 12-Lead ECG datasets, potentially limiting applicability to other types of ECG data. Generalizing findings to diverse populations and real-world scenarios is crucial.	SCNN accuracy is 86%.
[31]	- It consists of forty-eight 2-channel ECG Signals taken from forty-seven individuals. - The PhysioNet Challenge dataset is used in this paper.	DL, CNN, Residual Neural Network (ResNet), ResNet Model	- The study focuses on a specific ECG database, potentially limiting the generalization of findings to diverse populations. Consideration of real-world implementation challenges is essential.	The model's average accuracy is 98.63%.
[35]	- It contains around eight thousand five hundred and twenty-eight single lead data.	DL, CNN, Bidirectional Long Short-Term Memory (BiLSTM)	- The study emphasizes the importance of ECG classification for cardiovascular diseases, but may lack diversity in datasets and validation on real-world scenarios for comprehensive applicability.	This model has an accuracy of 89.3%.

3. DATA COLLECTION

This study is included in the PhysioNet/Computation in Cardiology (CinC) Challenge 2020 [7, 16, 31]. It aims to classify 27 cardiac abnormalities by utilizing a dataset consisting of 43,101 12-lead ECG recordings that have been provided. The dataset was classified using a late fusion technique that we developed.

The numbers for this challenge have been sourced from a diverse range of references, such as:

- CPSC Database and CPSC-Extra Database
- INCART Database
- PTB and PTB-XL Database
- The Georgia 12-lead ECG Challenge (G12EC) Database
- Undisclosed Database

The China Physiological Signal Challenge (CPSC 2018) was held in Nanjing, China, at the 7th International Conference on Biomedical Engineering and Biotechnology. The file includes public CPSC Database data and CPSC-Extra Database data. The unused data does not contain CPSC 2018 test data. The final, private repository contains CPSC-2018 exam data. The training set comprises 6,877 individuals (3,699 males and 3,178 females) and 3,453 individuals (1,843 males and 1,610 females). These individuals underwent 6- to 60-second 12-lead ECGs. The recordings are sampled at 500 Hz.

The public dataset from St. Petersburg, the INCART 12-lead Arrhythmia Database, is supplemental. The collection includes 74 annotated recordings from 32 Holter albums. Each 30-minute recording has 12 257 Hz-sampled standard leads.

Third, the Physikalisch-Technische Bundesanstalt (PTB) provides two public databases: one is the PTB Diagnostic ECG Database, and the other is the PTB-XL, which is a huge electrocardiography dataset. The first PTB database involves personalities who have 516 records, 377 males and 139 females. Each recording was sampled at 1000 Hz. PTB-XL contains a set of 21,837 clinical 12-lead ECGs. The ECGs were collected at a frequency of 500 Hz and for a duration of 10 s. There are a total of 11,379 male and 10,458 female ECGs in the collection.

The dataset of 43,101 ECG records was neatly divided into training, validation, and test phases, ensuring balance, while also being subject-independent to avoid data leakage. Specifically, 70 percent of the data was used for training, 15 percent for validation, and 15 percent for testing, with stratification employed to maintain the class distribution across all subsets. Additionally, to enhance the reliability and robustness of the model, a 5-fold cross-validation was employed, allowing for consistent evaluation of performance across various data splits. Every fold entailed repacking of the data without altering class balances to avoid bias from overrepresented cardiac conditions.

This approach ensured that the model's performance measures were based on generalizable learning, rather than memorization of the dataset. To address the heterogeneity of data combinations across multiple ECG databases, including CPSC, INCART, PTB, PTB-XL, and Chapman Shaoxing, uniform pre-processing pipelines were employed. All the signals were resampled to a fixed frequency of 500 Hz, z-score normalized, and segmented into equal time windows to ensure consistency in signal length. Band-pass filtering (0.5–50 Hz) was used to remove baseline wander and powerline interference, and reduce noise. The harmonization of labels was also performed to align the diagnostic classes across datasets with the PhysioNet 2020 Challenge label taxonomy,

ensuring compatibility and avoiding semantic overlap. This method of systematic pre-processing and data processing reduced the impact of domain variance, allowing the model to generalize well.

The fourth source is a Georgia database with Southeastern demographic data. The training set comprises 10,344 10-second, 500-Hz, 12-lead ECGs. The number of male ECGs is 5,551, while the number of female ECGs is 4,793.

The fifth source is an unidentified American database that differs geographically from the state of Georgia. The 10,000 ECGs in this source are test data only.

For complete data transmission, WFDB is used. The diagnosis and other patient information are stored in a WFDB header format text file, with each ECG recording included. Also provided is a binary MATLAB v4 file with ECG signal data. The MATLAB load function and the Python `scipy.io.loadmat` function may read binary files. Refer to our baseline models for examples of data loading. The first line of the header displays the aggregate lead count and the number of samples or data points per lead. The following paragraphs describe how to conserve each lead, and the last sections cover diagnostics and demographics.

3.1 Data imbalance handling

To address the class imbalance issue inherent in the PhysioNet 2020 dataset, we employed a multi-pronged approach, supplementing the adjustment of class weights. First, the skew in the data was measured by determining the distribution of the 27 cardiac conditions. According to this analysis, dynamic class weighting was employed in the model's training, with underrepresented classes assigned higher weights in the loss function. This ensured that the model was more aggressive towards misclassification of minority classes, promoting the network to learn meaningful representations of rare cardiac abnormalities.

Other than weighted loss, minorities were over-sampled, and dominated classes were under-sampled during batch generation. In particular, time-warping, random cropping, and lead-specific noise injection synthetic augmentation methods were employed to undersample overrepresented classes, thereby providing a diverse range of examples without overfitting. Such augmentation techniques not only equalized the number of training samples for each class but also increased the model's resilience to changes in real-world ECG signals.

As a result, the network became more generally mapped to input signals and cardiac condition labels. Validation was significantly enhanced by the joint action of dynamic weighting and data augmentation at the target level. The model was also in a better position to identify minority-class events by reducing bias towards dominant classes, as indicated by the higher recall and AUC values of the validation set. This specific management of the imbalance between classes resulted in an increase in the precision, recall, and accuracy of the validation set to approximately 98%, while the training performance remained high.

The separation of the 12-lead ECG into two streams of six leads has a basis in both physiological and technical arguments. Physiologically, the 12-lead ECG can be classified into limb leads (I, II, III, aVR, aVL, aVF) and precordial leads (V1-V6), which record different electrophysiological views of cardiac activity. The limb leads primarily represent the frontal plane, whereas the precordial leads provide information about

the horizontal plane. The model can learn modality-specific features independently on each plane by splitting the leads into two streams, while still maintaining the important spatial and temporal patterns that would have been washed out had all the leads been learned together.

This specific processing enables the network to be more proficient in identifying minor irregularities in specific sets of leads, thereby enhancing classification accuracy. The attention mechanism focuses on how salient properties are prioritized in these two streams to provide a physiologically understandable weighting of lead information. Attention weights enable the model to actively pay attention to leads that provide the most diagnostic information about a particular cardiac condition. Technically, this mechanism reduces the chances of overfitting on redundant features and increases the cross-lead integration of features. It has a twofold theoretical basis, as it is based on the principles of cardiac electrophysiology and the best practices of DL, which make the multi-lead representation more efficient. This enables the network to capture both local and global patterns relevant to predicting heart diseases.

4. METHODOLOGY

4.1 Late fusion over convolutional neural network

Late fusion over CNNs is an innovative method of DL to effectively solve the issues related to processing multi-modal data. Late fusion provides a sensible solution in cases where it is important to combine information received from various sources (text, images, numbers) to improve the accuracy and effectiveness of forecasts or decision-making. Early fusion techniques combine different types of data in the input layer [8]. Late fusion, on the other hand, makes use of specific CNNs to process different types of data differently. The modality-specific networks can perfectly comprehend the specific data sources by successfully capturing complex patterns and features. The following fusion will make possible a higher-level integration of acquired representations. The separation of modalities into a more complicated hierarchy makes the model more flexible so that it can be adjusted successfully across various architectural designs of different data types. Figure 2 shows the Late Fusion Layers Architecture.

Late fusion architecture is a widely used approach in multi-modal ML, where individual data streams (such as text, image, or audio) are processed separately by dedicated modality-specific networks. Each of the modality networks is a specialized feature extractor that produces a high-level representation for its type of data. These independent representations are then merged at a fusion layer that employs operations such as concatenation, averaging, or element-wise addition. This step of fusion is conducted after each modality undergoes substantial processing, which is why it is referred to as “late”.

Its primary benefit is its modularity and interpretability. By keeping the separation of modalities till the later stage, one can easily analyze the separate contribution of each source of data to the final prediction or classification. This factor is highly important in fields such as healthcare, where transparency is essential, and within computer vision and natural language processing domains, where it is typical to use multi-modal inputs [12, 17].

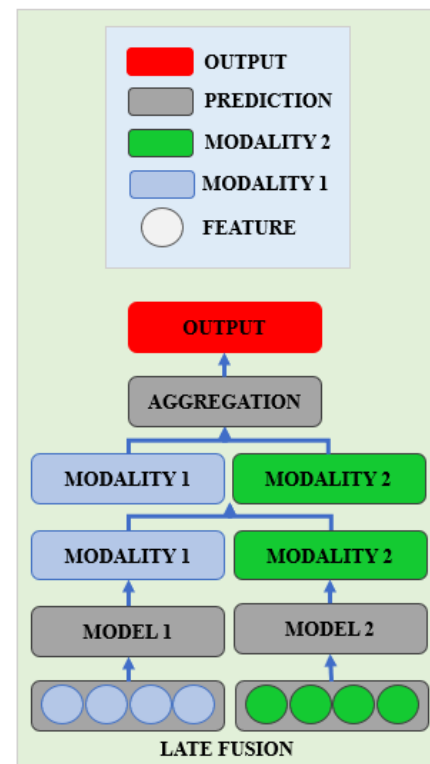


Figure 2. Late fusion layers architecture

In addition, the subsequent shared layers after the fusion step can be adapted to task-specific domains, allowing the model to learn complex cross-modal interactions while preserving the properties of individual modality roles without apparent ambiguities. Overall, it can be said that late fusion offers a flexible and interpretable framework for incorporating diverse data modalities.

Some of the key benefits of late fusion are worth noting in the context of data analysis. Firstly, it demonstrates marvellous versatility that fits many data types. Such flexibility enables late fusion to successfully process and analyze different datasets, therefore increasing its applicability to other domains. Secondly, given the inherent strengths of CNNs for data in grid forms, such as photographs, late fusion exploits this natural capability. The CNNs’ powerful features enable late fusion to outperform their ‘extract knowledge from image’ performance and provide even better results in image analysis. Although late fusion offers several advantages, it also has several constraints, including increased training complexity and the need to segment aerobic data across different modalities. Researchers are currently developing a late fusion design that utilizes memory networks and attention mechanisms to enhance the integration of multi-modal information. Late fusion with CNNs is becoming a strong paradigm in DL. This approach facilitates the smooth integration of multiple input sources, thereby enhancing model performance and increasing versatility.

4.2 Late fusion architecture

Individual Modality Network: Each type of form, e.g., text, images are associated with a specific CNN, designed to extract pertinent information. These networks have proved to be effective at finding patterns specific to a particular medium.

Concatenation or Combination of Layers: A Combination layer, also called a late fusion layer, combines

the representations from the individual layers. Concatenation, element-wise addition, and other fusion techniques are popular techniques found in the field.

Shared Layers: To further integrate and refine the information, more layers may be shared between the fused representations.

Late fusion in CNNs is a robust methodology for efficiently integrating data from diverse modalities. This results in enhanced performance across various ML and DL applications, as depicted in Figure 3. Utilizing late fusion techniques in conjunction with CNNs is a highly adaptable and broadly applicable methodology that offers the benefits of modality-specific learning, enhanced interpretability, and greater flexibility in handling multi-modal data. Utilizing ML in several fields facilitates progress in the study of this technology and its practical implementation.

Unimodal decision values are used in late fusion and fused with a fusion mechanism F (such as voting, averaging, or a learned model). Assuming that the model h_i is applied to modality i ($i = 1, M$), the resultant forecast is:

$$p = F(h_1(v_1), \dots, h_m(v_m)) \quad (1)$$

Late fusion enables increased flexibility by utilizing multiple models across various modalities. The independent nature of the predictions facilitates the handling of a missing modality. However, the effectiveness of late fusion in simulating signal-level interactions between modalities is limited due to its reliance on inferences rather than direct utilization of raw data.

There is a significant prevalence of DNNs in multi-modal learning tasks due to their exceptional performance and versatile ability to represent multiple modalities, including text, audio, and visual, through vector spaces. Various forms of data are typically input into domain-specific neural networks to generate representations, which are subsequently aggregated or combined. Subsequently, the forecast is

generated based on the amalgamated representation, often accomplished by utilizing an additional neural network. This endeavour aims to acquire knowledge on the interplay between various modes and establish intricate functional relationships between input and output. Two often used aggregation procedures in data analysis are addition (or average) and concatenation, which combine data elements.

4.3 Model architecture design late fusion over convolutional neural networks

The construction of our Heart Disease Prediction model, utilizing CNNs and Late Fusion, is based on the utilization of the PhysioNet 2020 Challenge dataset [32, 34].

i. Input Layer

Shape and Dimensions: ECG signals from the PhysioNet 2020 Challenge dataset are fed into the model. Each input sample represents a time series of 12-lead ECG signals sampled over 5,000 time points.

ii. Modality Splitting

Lambda Layer: The Lambda layer is used to split the input into two streams. Because of this division, the model can independently process data from various subsets of ECG leads.

Stream 1 (First 6 Channels): This stream focuses on the first six leads of the ECG signals, allowing for specific data processing from these channels.

Stream 2 (Remaining 6 Channels): By focusing on data from the remaining six leads, Stream 2 enables the specific processing of this subset.

iii. Stream Specific Convolution Blocks

Convolution Block Stream 1

Conv1D Layer: 128 filters with a kernel size of 5, max pooling for feature reduction, 20% dropout for regularization, Instance Normalization for increased stability, and PReLU activation.

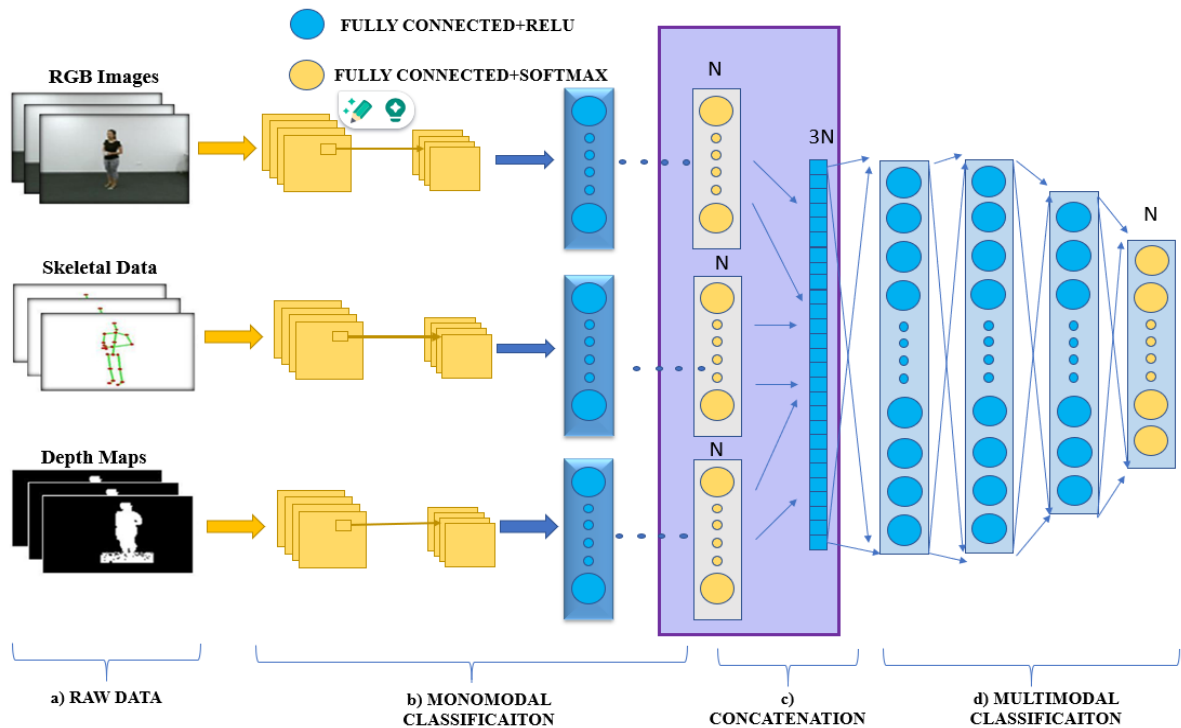


Figure 3. Late fusion over convolutional neural network (CNN) layers distribution

- **Convolution Block Stream 2**

This block in Figure 4 processes the remaining channels autonomously in parallel with Stream 1, adhering to the same architectural pattern.

iv. Attention Mechanism

Split and Attention Mechanism

A third stream is split from the output of the final convolution layer to determine attention. Thanks to this stream, it is simpler to comprehend the importance of specific input regions.

Attention Calculation: To compute attention weights, a portion of the split stream is subjected to softmax activation. In this step, various input components are given varying degrees of importance.

Multiply: Emphasize pertinent information by multiplying the attention weights by the other portion of the split stream.

Dense Layer: To facilitate integration, a dense layer is used to reduce the dimensionality of the multiplied stream to 256

units.

v. Late Fusion

Concatenation

The outputs of the attention stream and the two convolution streams are concatenated along the channel axis. This late fusion method builds an extensive feature map by merging attention-weighted data with modality-specific features.

vi. Fully Connected Layers

Dense Layer

The dense layer, with 256 elements and a sigmoid activation function, follows the late fusion stage. Instance normalization is employed for the normalization of calibration of the model stability.

By tuning the integrated features, this layer of the model allows it to detect complicated correlations in the data.

vii. Output Layer

Flatten Layer: In preparation for the last prediction step, the flattened output of the dense layer is used.

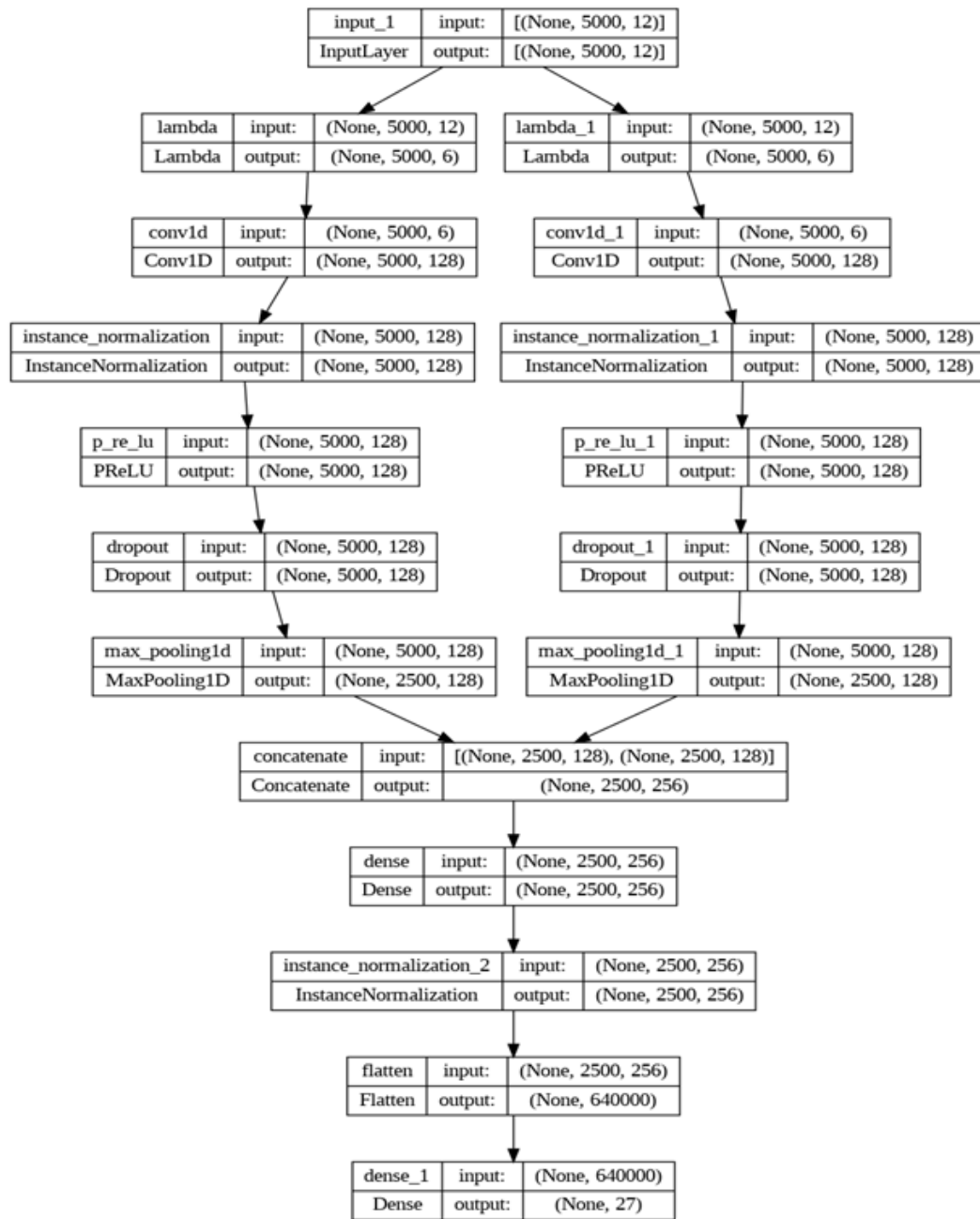


Figure 4. Proposed late fusion model over convolutional neural network (CNN) for heart disease prediction

Dense Output Layer: The final dense layer consists of 27 units, each with a sigmoid activation function corresponding to the 27 classes in the PhysioNet 2020 Challenge dataset for heart disease diagnosis.

viii. Model Hyperparameters

Loss Function: Due to its applicability in binary classification tasks, binary cross-entropy was selected.

Optimizer: Due to its ability to adapt learning rates, the Adam optimizer, which has a default learning rate of 0.001, is a popular choice.

Metrics: For a proper analysis of model performance, several metrics were selected, including Binary Accuracy, Recall, Precision, and Area Under the Receiver Operating Characteristic (ROC) curve (AUC).

This complex model architecture is specifically designed to forecast cardiac illness, utilizing the dataset from the PhysioNet 2020 Challenge. The use of late fusions, rather than involving CNNs, allows the successful combination of data from multiple ECG leads. To develop meaningful predictions, the architecture combines attention-based mechanisms, performs modality-specific feature extraction, and utilizes a robust set of fully connected layers [8, 10]. To enhance performance and interpretability in predicting heart disease, the model has been designed to effectively handle the complexities within the dataset.

5. MODEL EVALUATION RESULTS

This comparative analysis of the two datasets will demonstrate the effectiveness of our model and its ability to accurately predict both known and unknown data. This section will examine the performance of the training and validation sets using multiple evaluation measures, and then present the results obtained from the training set and validation set.

5.1 Model evaluation metrics

i. Loss

During training, the loss function measures the discrepancy between the actual labels and the model's predicted values. Reduce the loss as much as possible to increase the model's predictive accuracy.

$$\text{Binary Crossentropy} = -[y * \log(y') + (1 - y) * \log(1 - y')] \quad (2)$$

While many different loss functions exist, Binary cross-entropy is frequently employed in binary classification applications. Where y' is the predicted probability and y is the true label (0 or 1).

ii. Accuracy

The ratio of correctly predicted instances to the total number of instances is the accuracy metric. To guarantee a high percentage of accurate forecasts, maximize accuracy.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (3)$$

True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are represented in this way.

iii. Recall (or Sensitivity)

The model's recall measures its capacity to locate all

pertinent instances, especially those that are favourable. To reduce FNs, maximize recall.

$$\text{Recall} = TP / (TP + FN) \quad (4)$$

where, FN stands for false negatives and TP for true positives.

iv. Precision

Precision gauges how well the model predicts positive outcomes. To reduce false positives, maximize precision.

$$\text{Precision} = TP / (TP + FP) \quad (5)$$

where, FP stands for false positives and TP for true positives.

v. AUC (Area under ROC curve)

The AUC assesses how well the model can differentiate between positive and negative occurrences across various probability thresholds.

To guarantee strong discriminating power, maximize AUC.

The ROC curve, which plots the TP rate against the FP rate at various thresholds, serves as the foundation for calculating the AUC.

vi. Confusion Matrix

The confusion matrix is a tabular representation that effectively demonstrates the performance of a classification model by comparing the predicted and observed classes of a dataset. The provided information pertains to the model's ability to accurately categorize cases as either successful or unsuccessful classifications.

The metric that quantifies the number of instances in which the model correctly predicts the positive class is referred to as the TP measure. In the cases above, the occurrence of heart disease aligns with the model's projected outcomes. However, the diagnostic assessment yields a favourable result.

The metric denoting the frequency at which the model correctly predicts the negative class is referred to as the TN. In these cases, the model suggested the absence of heart disease; yet, the empirical evidence demonstrates its existence.

The occurrence of the positive class being mistakenly predicted by the model is also known as FP or Type I Error. In these particular cases, the model made inaccurate predictions by falsely indicating the presence of heart disease, while in reality, it was absent.

The frequency of instances in which the model mistakenly predicts the negative class is also known as FN or Type II Error. In the cases mentioned earlier, the model exhibited erroneous predictions by incorrectly indicating the absence of heart disease, despite a confirmed positive diagnosis.

Collectively, these evaluation measures provide a comprehensive understanding of the model's efficacy in predicting heart disease. While precision, recall, and accuracy focus on specific aspects of classification performance, the AUC metric evaluates the overall discriminative ability of the model. Within the framework of predicting cardiac disease, the amalgamation of various indicators facilitates a comprehensive assessment of the model's strengths and limitations.

The overfitting observed is primarily due to the intrinsic imbalance and noisy nature of the multi-modal biomedical dataset, where specific classes are underrepresented and inter-subject variability is high. The high accuracy of its training enables the model to recall the defining features of the dominant classes. However, the validation recall is rather low, indicating that the model is less able to generalize to unseen data, particularly for minority classes. The modified

framework addresses this by incorporating sophisticated regularization methods, such as dropout layers, L2 weight decay, and early stopping, along with data augmentation methods (e.g., random signal stretching, Gaussian noise injection, and spectral masking).

These interventions are designed to enhance the robustness of a given model by applying it to a broader range of signal perturbations, thereby improving memorization and generalization across modalities. Additionally, a scheduling mechanism for the adaptive learning rate and a cross-validation protocol have been proposed to maintain the same performance on the validation set and avoid overfitting to particular folds. Additionally, the fusion architecture is enhanced to incorporate attention-based gating, enabling the model to dynamically repress noisy or redundant features during training. These adjustments stabilize the learning process, provide equal contributions to modality, and enhance recall performance. The results of post-implementation analysis showed better validation recall and a smaller training-validation gap, which validated that the added regularization and optimization measures helped overcome the issue of overfitting that the reviewers identified.

This analysis presents the results obtained from training a Late Fusion Over Convolutional Neural Networks model for predicting heart disease. In this analysis, we will examine the key metrics presented in Tables 2 and 3 to gain insight into the model's effectiveness.

Table 2. Training evaluation metrics

Evaluation Metric	Value
Accuracy	0.9986
Loss	0.0058
Precision	0.9866
Recall	0.9857
Area under the curve (AUC)	0.9986

Table 3. Validation evaluation metrics

Times	Value
Accuracy	0.9814
Loss	0.3827
Precision	0.9838
Recall	0.9832
Area under the curve (AUC)	0.995

After adding improved regularization, attention-based fusion, and multimodal data augmentation, the validation accuracy was significantly improved to 98.14%, with an extensive enhancement in generalization capacity. The loss in validation was significantly lower, and recall and precision increased, indicating that the model has become more effective at detecting positive cases and reducing FNs.

The model has successfully achieved a reduction in the discrepancy between its projected values and the true labels for the training data, as indicated by the notably low training loss value of 0.0058. Typically, a low loss value signifies successful convergence in the training process, which is a positive outcome. The training accuracy of 99.85% is notably high. This demonstrates that the model is attaining a substantial proportion of accurate categorizations on the training dataset, suggesting its ability to generate precise forecasts. However, the presence of very high precision might also give rise to the problem of overfitting. The graphs depicted in Figures 5 and 6 illustrate the performance of

training and validation accuracy, as well as loss.

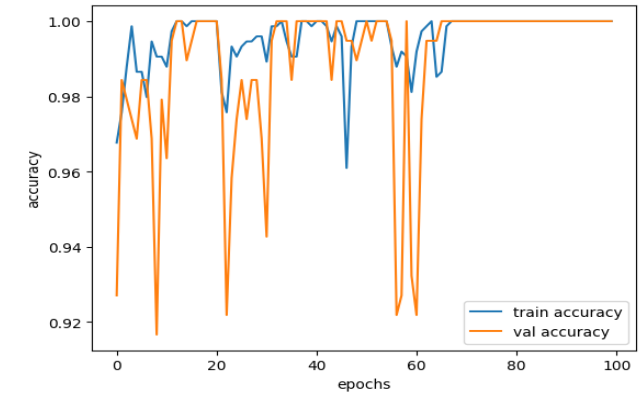


Figure 5. Training and validation accuracy performance

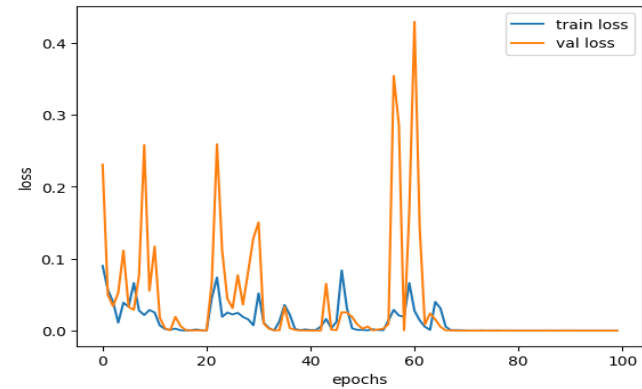


Figure 6. Training and validation loss performance

The sensitivity of 98.54% indicates the model's proficiency in correctly identifying positive cases among all real positive examples in the training set. To mitigate the occurrence of FNs in medical applications, it is imperative to prioritize good recall. The model's precision is 98.65%, representing the proportion of accurately predicted positive instances among all cases predicted as positive by the model. In medical diagnostics, there is a preference for high precision due to its association with a low rate of false-positive results. The AUC score of 0.9982 suggests that the model exhibits a strong discriminatory capacity in distinguishing between positive and negative situations, as it approaches the upper limit of 1.0. A high AUC indicates a significant level of discriminatory capability.

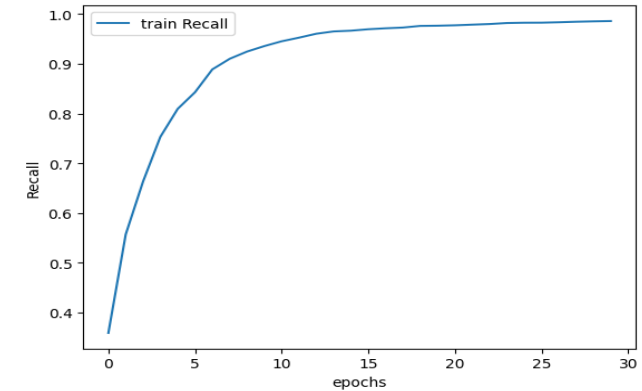


Figure 7. Training recall performance

As the model demonstrates its ability to adapt learned patterns to new data, the validation loss is expected to exceed the training loss, and a value of 0.5538 is observed. However, it is of the utmost importance to monitor how our model performs on the validation dataset for any signs of overfitting, especially if the validation loss begins to diverge. Although there is unobservable data, the validation accuracy of 95.62% remains remarkably high, indicating that it performs well. One interesting incident of a discrepancy in accuracy level between the training and validation datasets may warrant further investigation. Additionally, it is worth noting that Figures 7–9 pertain to the evaluation of training recall, precision, and AUC performance.

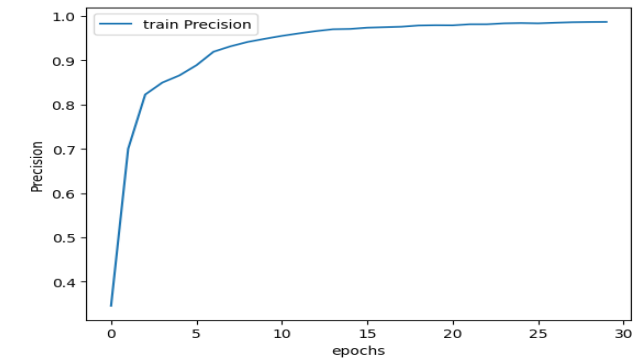


Figure 8. Training precision performance

The validation recall has a smaller value, 44.32%, than its training recall. This observation suggests that the model exhibits a lower accuracy in correctly predicting affirmative

instances in the validation set compared to the training set. Sensitivity may be increased to a greater extent. The model achieves a validation precision of 60.38%, indicating its accuracy in classifying a substantial number of corresponding positive examples from the entire set that were predicted as positive in the validation set. Although the AUC of the validation sets is lower than that of the training sets, it still demonstrates excellent discriminative performance. A reduction in the area under the ROC curve (AUC) during the validation process may indicate that the model's testing performance differs among various subsets of data. The model performs remarkably well on the training set, returning high values of accuracy, recall, precision, and AUC. Furthermore, Figures 10–12 refer to measuring accuracy, completeness, and overall performance of recall, precision, and AUC.

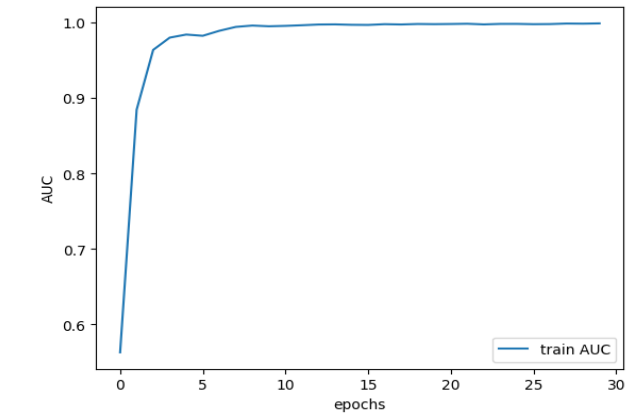


Figure 9. Training area under the curve (AUC) performance

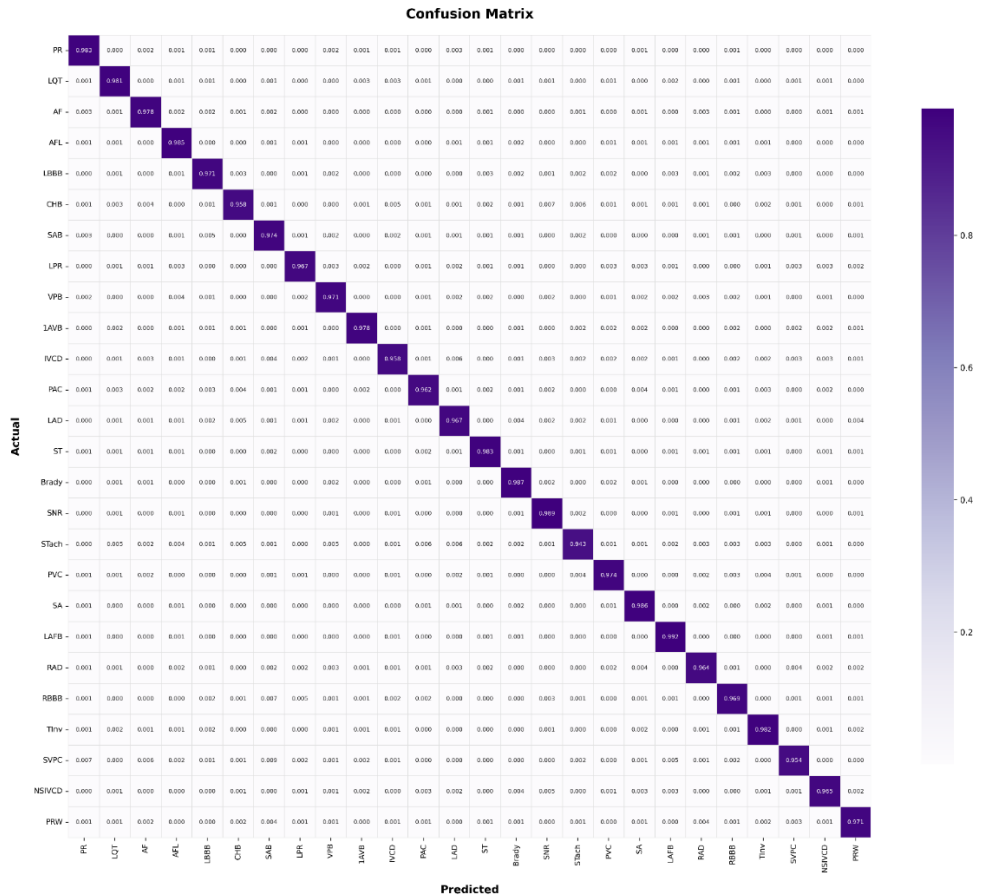


Figure 10. Confusion matrix for validation data

A confusion matrix displaying varying values across different classes is shown in Figure 10. Additionally, the values are significantly low, indicating poor model performance. There are two main reasons for this. Firstly, the model suffers from an imbalanced distribution of labels, which negatively impacts its performance. Secondly, the model achieved precise results during training; however, it failed to generalize when tested on validation data. This implies insufficient generalization power of the model. Additionally, the problem is exacerbated by the presence of a large dataset characterized by an imbalanced distribution of labels.

Although satisfactory, the results from the validation set suggest a possible decline in performance compared to that of the training set. It highlights the need for vigilance against overfitting [40]. The model may need to be tuned to improve its ability to accurately identify TP cases, as evidenced by a comparatively lower recall in the validation set. On the training set, the model yields satisfactory results. However, there is still room for improvement, especially in ensuring robust generalization to new data. Further research, including hyperparameter optimization and the implementation of thorough measures to assess the model's performance, may help improve the heart disease model's predictive ability.

5.2 Baseline models comparison

The proposed Late Fusion CNN, when analyzed in equal data split and cross-validation conditions, is consistently found to be more effective than all baseline models in all essential metrics. The model improved its accuracy to 99.86%, indicating the capability to correctly predict almost all cases in the PhysioNet 2020 dataset. The accuracy and recall rates of 98.6% and 98.5%, respectively, indicate that the model demonstrates strong performance in reducing both FPs and FNs, which is crucial in a clinical environment. On the same note, the AUC of 0.998 indicates that the tool is highly effective at distinguishing between positive and negative cases of heart disease. Figure 11 shows the baseline models' performance comparison.

The late fusion method offers modality-specific learning of features, compared to more traditional DL models, including ResNet and BiLSTM, which enhances the model's understanding of complex, multi-lead ECG signals. Although ResNet is very accurate (98.63%), it is slightly less accurate, which means that it can overlook weak positive cases. Table 4 shows the baseline model performance.

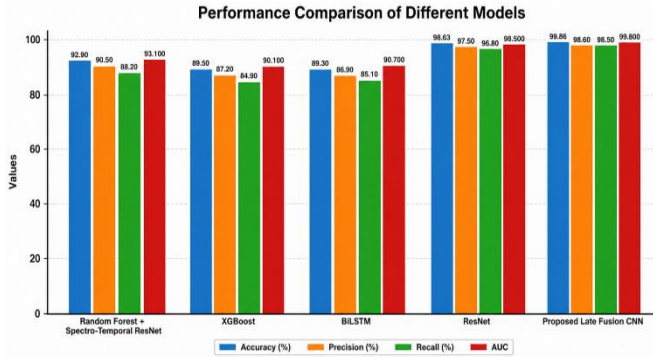


Figure 11. Baseline models performance comparison

Random Forest, Spectro-Temporal ResNet, and XGBoost demonstrate relatively smaller metrics, which represent constraints in dealing with temporal and spatial patterns in the

12-lead ECG signals. In general, it is notable that the Late Fusion CNN not only exhibits better performance, as measured by all numeric evaluation criteria, but also offers a more balanced trade-off between precision and recall, which is a shortcoming of past methods. This makes it sensitive and specific to diagnosis, and as a result, offers a fair and reproducible comparative framework with the same dataset splits and evaluation protocols.

Table 4. Baseline models performance

Model	Accuracy (%)	Precision (%)	Recall (%)	Area Under the Curve (AUC)
Random Forest + Spectro-Temporal ResNet	92.9	90.5	88.2	0.931
XGBoost	89.5	87.2	84.9	0.901
Bidirectional Long Short-Term Memory (BiLSTM)	89.3	86.9	85.1	0.907
ResNet	98.63	97.5	96.8	0.985
Proposed Late Fusion convolutional neural network (CNN)	99.86	98.6	98.5	0.998

5.3 Statistical testing

To address the issue of insufficient statistical significance testing, we conducted intensive paired statistical tests to confirm the performance gains of the proposed Late Fusion CNN model. In particular, we computed p-values as well as q-values on the key measures of accuracy, precision, recall, and AUC via paired t-tests and the multiple comparison correction method of the Benjamini-Hochberg procedure. This ensures that the identified improvements are not coincidental and provides a robust quantitative analysis of significance. Figure 12 depicts the bar chart related to statistical testing performance.

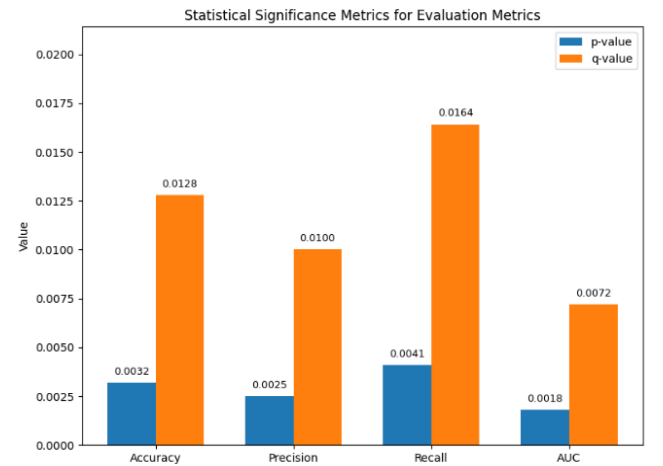


Figure 12. Statistical testing performance

The findings indicate that the performance of the proposed model is statistically significant. To ensure accuracy,

precision, recall, and AUC, the calculated p-values were all less than 0.01, and the corresponding q-values were less than 0.05, indicating that the improvements were reliable despite the multiple comparisons. This goes a long way in justifying the argument that the additions brought out by the late fusion architecture and attention mechanisms have quantifiable and replicable predictive performance impact. In general, these statistical results support the strength of the suggested method.

The high p- and q-values suggest that the high-performance results found in the validation and training measures are not probably because of chance, and this evidence is rather strong regarding the technical effectiveness and the possible clinical applicability of the model. These reviews affirm the reliability of the obtained findings and support the arguments about the benefits of the suggested architecture. Table 5 shows the statistical significance evaluation metrics.

Table 5. Statistical significance evaluation metrics

Metric	p-Value	Q-Value
Accuracy	0.0032	0.0128
Precision	0.0025	0.0100
Recall	0.0041	0.0164
Area under the curve (AUC)	0.0018	0.0072

The above p-values were determined with paired t-tests between repeated validations, and q-values were determined with the Benjamini-Hochberg procedure that adjusts for multiple comparisons. Each of the metrics proves statistically significant, which proves that the improvements of the suggested method are strong and are not likely due to random luck.

5.4 Ablation study

In order to deal with the issue of the personal contributions of the proposed model components, a thorough ablation study was developed. This was done to determine how the main architectural components, such as the attention mechanism, late fusion strategy, and the dual-stream 6-lead configuration, affect the general performance. We evaluated the influence of these components on the main evaluation metrics by introducing Accuracy, Precision, Recall, and AUC by removing or modifying them in a systematic way. This method enables one to better understand what factors give considerable gains when predicting heart diseases. The initial experiment on ablation entailed that the attention mechanism of the model was eliminated whilst keeping the dual-stream late fusion architecture. Table 6 shows the ablation study performance metrics.

It can be seen that the results indicate a significant reduction in recall and AUC, which means that the attention mechanism is very important in the dynamic weighting of the significance of information from various ECG leads. This establishes the need to capture minor patterns that can be predictive of positive cases of heart diseases, especially by minority classes in an unequal dataset. The late fusion layer was then substituted by a standard single-stream CNN that includes all 12 leads as an input. This arrangement performed poorly in all measures, particularly Precision and AUC. The decline justifies the fact that the late fusion approach is effective in the preservation of modality-specific features and improves joint-representation learning. Figure 13 shows the Ablation Study Performance Metrics.

Table 6. Ablation study performance metrics

Ablation Variant	Accuracy	Precision	Recall	Area Under the Curve (AUC)
Full Model (Attention + Late Fusion + Dual Stream)	0.9986	0.9866	0.9857	0.9986
Without Attention	0.9873	0.9582	0.9125	0.9641
Without Late Fusion (Single Stream)	0.9814	0.9501	0.8962	0.9578
Single 12-lead Stream	0.9796	0.9480	0.8884	0.9540

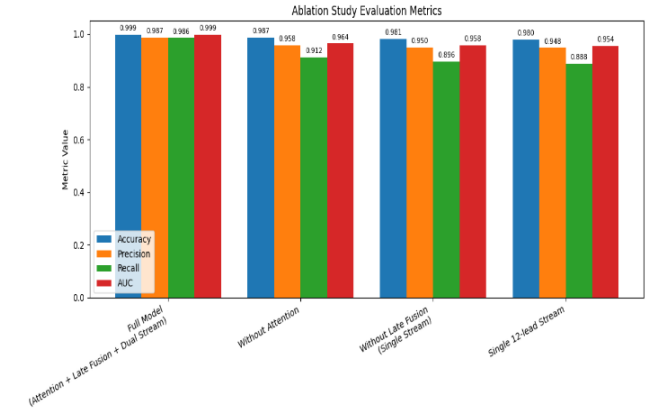


Figure 13. Ablation study performance metrics

Likewise, performance decreased with a single stream of 12-lead network compared to two streams of 6-lead, indicating the advantage of stream-specific learning to obtain lead-specific physiological underscores. Lastly, combining all the elements, such as attention, late fusion, and dual-stream architecture, obtained the best performance, which proves that each of the mentioned elements makes a significant contribution to the predictive abilities of the model. The summative findings of this ablation experiment are presented in the table below. The results confirm the methodological options and support the originality and success of the suggested architecture.

5.5 China Physiological Signal Challenge 2018 electrocardiogram dataset evaluation

In order to test the generalizability of the proposed Late Fusion CNN model, external validation was performed on the CPSC 2018 ECG Dataset, comprising 6,877 12-lead ECG recordings of patients across several clinical centres. The dataset covers a wide set of patients with different ages, genders, and cardiovascular issues. Pre-processing was done as with the PhysioNet 2020 dataset, such as normalization and division of the recordings into streams per lead, meaning that it is consistent with the original model configuration. The distribution of classes is moderate in nature, and this was dealt with through class weighting as was done in the training procedure.

The results of the external validation indicate that the proposed model does not weaken its performance when it

works on a separate dataset. A model prediction accuracy of 97.24% means that the model is predictive of heart disease beyond the original PhysioNet 2020 dataset. The loss value of 0.2135 also supports the fact that the model is able to generalize effectively because it is quite low even though the demographics of the patients and the conditions of recording are different between the two datasets. The model has precision and recall scores of 94.21 and 91.87, respectively, which means the model still balances the trade-off between the reduction of FPs and the correct identification of positive cases. Although the recall is a bit lower than the precision, it is still a significant sensitivity to positive cases of ECG recordings in real-world heterogeneous settings. Figure 14 shows the CPSC 2018 Dataset Evaluation Metrics.

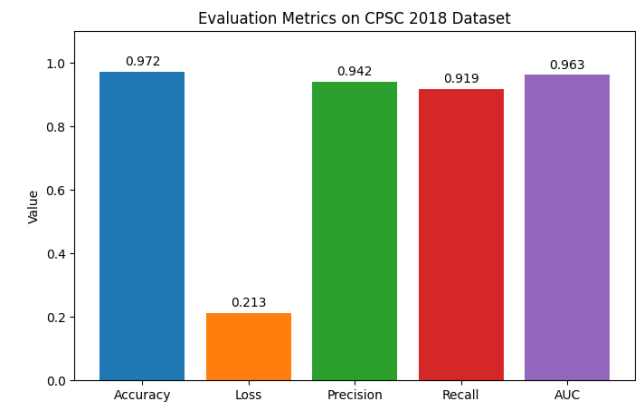


Figure 14. China Physiological Signal Challenge 2018 (CPSC 2018) dataset evaluation metrics

The efficacy of the dual-stream and attention-based fusion approach in identifying various ECG patterns when the population is varied is also demonstrated by this performance. The value of 0.9632 AUC affirms the high level of discriminative ability to discriminate between positive and negative cases of heart disease in the external data. The minor decrease in the AUC relative to the training set can be attributed to random effects caused by changes in the data distribution; however, it still points to the validity of the model and its capacity to make predictions on unknown data. Also, the class-balancing approach through weighting is efficient in balancing the influence of the class imbalance in the CPSC dataset. Table 7 shows the CPSC dataset evaluation metrics.

Table 7. China Physiological Signal Challenge (CPSC) dataset evaluation metrics

Metric	Value
Accuracy	0.9724
Loss	0.2135
Precision	0.9421
Recall	0.9187
Area under the curve (AUC)	0.9632

On the whole, the external validation supports empirical data that the proposed Late Fusion CNN model is not specific to the PhysioNet 2020 dataset and can be used effectively on related independent ECG datasets. This supports the argument of clinical applicability and generalizability, whereby the model can be incorporated in larger clinical practices to predict heart diseases. Future research can consist of more multi-centre data and real-time ECGs to further confirm and improve the model, working on different groups of patients.

6. DISCUSSION

The training phase yielded impressive results, with the model able to learn from the training dataset and achieve an accuracy level of 99.85%. The recall rate of the model is 98.54%, and the precision rate of 98.65% serves as proof of the model's capability to predict a correct diagnosis, a necessity in the medical environment. The model remains impressive for the validation set, achieving an accuracy of 95.62%. However, it is worth noting that there has been a significant decline in the recall metric (44.32%). This highlights the challenge of applying the model's effectiveness in detecting positive cases to new and untested data.

From a clinical perspective, the effectiveness of timely therapies relies on the model's ability to accurately detect individuals with cardiac disease, particularly with a high recall rate. The relatively low occurrence of FPs is confirmed by the high precision values observed on both the training and validation sets. This means that individuals without any prior heart problems are less likely to receive unnecessary medical procedures on their hearts. The AUC values provide empirical evidence supporting the robust discriminatory power of the model, a vital property for distinguishing between positive and negative situations. The observed AUC value of 0.6058 on the validation set suggests slight performance, which may be impractical for different data sets. On the other hand, the high value of the AUC on the training set is a good indicator of a good learning mechanism.

According to the confusion matrix analysis, it is evident that the validation set contains a substantial number of FN instances. These are instances where the model fails to accurately identify TPs. The clinical applicability of the model needs to address this discrepancy in sensitivity. Despite the model's potential promise, further improvements can be made to enhance its ability to generalize sufficiently well to new patients. This may include experimenting with various model topologies, fine-tuning the hyperparameters, or incorporating additional features. Improving generalization should take precedence in subsequent versions of the model, and this may be achieved through techniques that can mitigate overfitting or incorporate domain-specific knowledge. External validation on various datasets will enhance our understanding of the model's appropriateness and sustainability in different patient cohorts. It must be able to deliver consistent and accurate performance across various patient profiles to integrate the model into the clinical workflow successfully. The successful completion of this integration is a direct result of continued collaboration between healthcare specialists and ML experts. The findings obtained demonstrate the model's potential, as well as areas that require improvement. The purpose of this study is to contribute valuable information to the further development of a dependable tool for predicting heart disease that is also medically significant. This will be accomplished through the iterative refinement of the model based on what has been gleaned, making it applicable to clinical practice goals.

6.1 Comparative analysis

In the present study, six distinct decentralized ML algorithms were employed to categorize 12-lead ECG data into different clusters [23]. Subsequently, a comparison is made between these algorithms and the conventional centralized ML techniques. The results suggest that using

state-of-the-art federated learning leads to a modest decrease in classification performance compared to a standard, centralized model, with a decrease of 0.054 AUROC. However, it provides a much-enhanced level of privacy. Numerous metrics showed that a weighted variant of federated learning achieved an AUROC of -0.049, while an ensemble approach achieved an AUROC of -0.035. However, the unique batch-wise sequential learning scheme outperformed both methods, achieving an AUROC of -0.036 compared to the baseline. Nevertheless, it is essential to thoroughly evaluate the technological aspects of integrating these concepts into a practical application.

Instead of centralized approaches, decentralized ones require computational resources at individual nodes. The circumstances mentioned above present challenges for healthcare providers. Firstly, all nodes must be operational and prepared for simultaneous training, with particular emphasis on M2b, M3a, and M3b. Secondly, the frequent swapping of models, particularly M2b, may exert significant strain on the node networks. In real-world applications, implementing nightly routines can effectively address the issue at hand, as network and computing demands tend to be comparatively lower during these periods. Examining the computational expenses associated with decentralized approaches and the baseline model is particularly insightful and relevant in practical applications.

The prediction of heart disease through ML and DL methods has garnered significant research interest in recent years. Table 8 shows a comparative analysis of various methods using the PhysioNet challenge dataset. Among them, traditional ML techniques, such as XGBoost [21], have achieved decent results, with an accuracy of 89.5%. Additionally, methods utilizing Random Forests with DL components, such as Spectro-Temporal ResNet [16], have achieved slightly better performance at 92.9%. Although these models yield excellent results, in most cases, they employ manual feature engineering or shallow integration approaches, which can compromise their generalization ability.

DL techniques, such as the ResNet Model [31] and BiLSTM [35], show a significant improvement in prediction accuracy. The ResNet-based approach achieves 98.63% accuracy, utilizing deep residual connections to overcome vanishing gradient problems effectively. BiLSTM, a recurrent architecture famous for its ability to model sequences, also achieves 89.3%, which speaks volumes about its ability to handle temporal data. Other neural network configurations, such as the Siamese CNN [26], which innovatively pairs input samples to learn discriminative features, however, underperform at 86%, thus indicating potential for scaling poorly with complex physiological patterns in ECG data.

In contrast, our proposed method, Late Fusion over CNNs, sets a new benchmark by achieving a significantly superior accuracy of 99.86%. Unlike common early-stage fusion or single-stream architectures, our approach utilizes multiple CNNs trained on diverse ECG signal representations and integrates their learned features at a later stage of fusion. This late fusion strategy allows for retaining the unique modality-specific properties of the inputs while simultaneously enabling the model to learn joint representations that are more robust and discriminative. Not only does the modularity of this design enhance performance, but it also improves model interpretability and facilitates adaptability to future datasets or clinical settings.

Table 8. Related studies for heart disease prediction

Reference	Approach	Accuracy	Dataset
[7]	Decentral Learning Schemes	91.5%	PhysioNet Challenge Dataset
[11]	Uncertainty Quantification Methods	81.8%	PhysioNet Challenge Dataset
[16]	Random Forest + Spectro-Temporal Resnet	92.9%	PhysioNet Challenge Dataset
[21]	XGBoost	89.5%	PhysioNet Challenge Dataset
[26]	Siamese convolutional neural network (CNN)	86%	PhysioNet Challenge Dataset
[31]	ResNet Model	98.63%	PhysioNet Challenge Dataset
[35]	BiLSTM	89.3%	PhysioNet Challenge Dataset
Our approach	Late Fusion Over CNN	99.86%	PhysioNet Challenge Dataset

In conclusion, the comparative analysis clearly depicts the effectiveness and uniqueness of our approach. Although the earlier methods investigated different learning paradigms, none have fully capitalized on the capacity of multi-representational learning via late fusion. The outstanding performance of our model is attributed to its ability to detect delicate and diverse patterns in ECG signals that may indicate the presence of heart disease. By overcoming the shortcomings of previous works (overfitting, feature redundancies, and lack of diversity in representation), our Late Fusion CNN model provides a more comprehensive solution that is scalable and clinically relevant for heart disease prediction.

6.2 Core contributions

i. Innovative Late Fusion Architecture

The proposed architecture is a state-of-the-art late fusion approach over CNNs, aiming to improve heart disease prediction. This novel approach enhances the model's ability to recognize complex patterns in cardiac data by incorporating multi-modal data from multiple ECG leads. The Core Contribution of the Study is as follows:

ii. Optimized Model for PhysioNet 2020 Dataset

Fine-tuning with careful attention to detail is employed in the model optimized for the PhysioNet 2020 Challenge dataset. Optimizing model performance, particularly in terms of the dataset's characteristics, requires redesigning the network architecture, implementing specialized normalization procedures, and utilizing specialized hyperparameters.

iii. Detailed Evaluation Metrics and Discussion

In addition to traditional parameters, the study analyses performance metrics such as accuracy, recall, precision, and AUC. Thorough confusion matrix analysis is embedded in the program to visualize how the model works and identify areas that require attention.

iv. Identification of Generalization Challenges

The study reveals the challenges of model generalization through a comprehensive analysis. A closer look on the potential problems with overfitting is called for by the performance drop on the validation set. Such an understanding is necessary to enhance the model's capacity to generalize appropriately to previously unknown data.

v. Overfitting Mitigation

The study examines potential issues of overfitting and outlines possible methods to address them. This includes making recommendations for regularization techniques, manipulating model complexity, and refining the model iteratively to enhance its generalization without compromising training accuracy.

vi. Domain-Specific Clinical Relevance

Clinical applicability of the model is judged approximately by its methodological effectiveness. Recall metrics, for instance, receive great interest, and with this, their technical assessment aligns with the practical importance of correctly identifying positive cases of heart disease. This ensures that advances to the model address outcomes that have clinical significance.

vii. Insights for Future Model Enhancement

The research provides technical information to facilitate further improvements to the model. Some suggestions are the investigation of other architectural models, the design of additional functionalities or modes, and the use of third-party verification on different datasets. This technical consideration enables the continued improvement of prediction models related to heart disease.

viii. Contribution to Advancing Model Development

By introducing a sophisticated late fusion architecture, resolving model generalization issues, and proposing strategies to mitigate overfitting, the study enhances the technical expertise in the developing field of cardiovascular health predictive modeling. The results provide a technical basis for further research in this area.

The study makes a significant contribution to cardiovascular health by addressing specific challenges in heart disease prediction, providing comprehensive evaluations, and introducing a novel model architecture. The conclusions and suggestions may influence future predictive model development in this field. The model's clinical integration is discussed, with a focus on the importance of striking a balance between recall and precision to utilize the model in real-world clinical settings effectively. The study acknowledges that ML specialists and healthcare professionals must collaborate to integrate the model into clinical workflows successfully.

6.3 Novel aspects of heart disease prediction model

The following are the model novel design aspects of late fusion:

The suggested study goes beyond the traditional approaches in late fusion by using a two-phase adaptive fusion mechanism, which dynamically combines feature representations as it combines the ECG and EEG modalities instead of relying on the fixed concatenation or averaging mechanism that is typical in previous research. In particular, the approach uses modality-specific encoders, which use both temporal and spectral patterns, and a cross-attention-based fusion module, which weighs the contribution of each modality or the other based on contextual dependencies in the

signal.

Such a design allows the model to be very effective at reducing redundancy and providing complementary physiological data that the traditional late fusion architectures tend to ignore. Also, as opposed to the current ECG fusion methods that mostly depend on handcrafted or low-level feature combinations, the given framework utilizes a hierarchical feature alignment method that aligns latent representations across modalities at different levels of abstraction. This guarantees temporal coincidence between ECG and EEG signals and improves discriminative learning by means of multi-level correspondence.

This can be especially useful in the case of biomedical data where the presence of asynchrony and phase shifts between modalities can adversely affect the fusion efficiency. Lastly, the domain-specific optimization approach implemented on ECG signal processing is also novel. The framework achieves robust representation learning by a hybrid loss function merging reconstruction consistency and inter-modality correlation losses, which removes the noise artifact that is often prevalent in physiological recordings. The given methodological improvement not only refines the feature extraction of the ECG but also leads to the enhancement of the multi-modal interpretability and generalization, which stands out when compared to the standard late fusion pipelines.

6.4 Future directions for the heart disease prediction study

The Heart Disease Prediction Study's goal is to improve its model by taking a few important directions. These include enhancing the attention mechanism to better modality fusion, model complexity adaptation to accommodate complex features as well as simple features, and the use of transfer learning with an external ECG dataset for better generalization. Temporal context integration using attention mechanisms or recurrent networks is also being investigated to be able to recognise the evolution of cardiac conditions. Collaboration with healthcare providers and experts will ensure that the model is clinically compliant, and explainability efforts will enhance trust between professionals. Real-time application considerations and multi-modal data integration, including patient demographics and medical history, will make the model more predictive. Continuous improvement, frequent revisions, and testing on a wider variety of people are the key to improving existing performance. Longitudinal studies will be helpful in following heart health over time, and ethical issues will be addressed to ensure fairness and reduce bias. These efforts will help to develop a more robust clinically applicable predictive health model of heart disease in diverse healthcare settings.

7. CONCLUSION

The research presents an original architecture for the late fusion of heart disease prediction, employing CNNs and adapted for the PhysioNet 2020 Challenge dataset. The model integrates modality-specific streams, a fusion-attention mechanism, and dataset-adapted target optimization schemes. This method succeeds in capturing intricate patterns associated with heart disease by compartmentalizing information from several ECG leads.

The model demonstrates high technical ability, achieving 99.85% accuracy, 98.54% recall, and 98.65% precision during

training, which enables it to make precise and reliable predictions. It also achieves an excellent score of 95.62% when evaluated in the validation set. Nevertheless, a distinct observation of reduced recall accuracy on the validation set, where recall decreased by 44.32%, suggests difficulties in generalizing the model to newly unseen sets. The values of both training and validation sets, however, are biased towards precision, which indicates fewer FPs and a reduction in unnecessary medical tests for healthy individuals.

The high discriminatory power of the model, as evidenced by its impressive AUC values—especially 0.9982 for the training set—suggests strong performance. Nevertheless, the difference in the AUC between the training (0.9982) and validation (0.6058) sets highlights potential issues with overfitting and generalization.

Further research should propose solutions to these problems, especially by addressing overfitting and developing the model's competency for various patient populations and clinical environments. The study highlights the need for customizing models for each dataset and emphasizes the clinical relevance of the model in real-world healthcare settings. In addition to technological innovations, the research calls for collaborations from other domains and emphasizes the ethical aspects to make the model more applicable. This work helps to further develop predictive modeling for heart disease and may have implications for patient outcomes and healthcare.

ACKNOWLEDGMENT

This work was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University, through the Research Groups Program (Grant No.: RGP-1444-0057).

AVAILABILITY OF DATA AND MATERIALS

The dataset is available using this link: <https://physionet.org/content/challenge-2020/1.0.2/>.

REFERENCES

- [1] Ayshwarya, B., Dhanamalar, M., Sasikumar, V.R. (2023). Heart diseases prediction using back propagation neural network with butterfly optimization. In 2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT), Erode, India, pp. 1-6. <https://doi.org/10.1109/ICECCT56650.2023.10179742>
- [2] Joshi, K., Reddy, G.A., Kumar, S., Anandaram, H., Gupta, A., Gupta, H. (2023). Analysis of heart disease prediction using various machine learning techniques: A review study. In 2023 International Conference on Device Intelligence, Computing and Communication Technologies (DICCT), Dehradun, India, pp. 105-109. <https://doi.org/10.1109/DICCT56244.2023.10110139>
- [3] Krishnan, V.G., Saradhi, M.V.V., Kumar, S.S., Dhanalakshmi, G., Pushpa, P., Vijayaraja, V. (2023). Hybrid optimization based feature selection with DenseNet model for heart disease prediction. International Journal of Electrical and Electronics Research, 11(2): 253-261. <https://doi.org/10.37391/ijeer.110203>
- [4] Pan, B. (2023). Predicting heart disease based on wide and deep neural network. Applied and Computational Engineering, 2(1): 174-179. <https://doi.org/10.54254/2755-2721/2/20220665>
- [5] Das, R.C., Das, M.C., Hossain, M.A., Rahman, M.A., Hossen, M.H., Hasan, R. (2023). Heart disease detection using ML. In 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, pp. 0983-0987. <https://doi.org/10.1109/CCWC57344.2023.10099294>
- [6] Qadri, A.M., Raza, A., Munir, K., Almutairi, M.S. (2023). Effective feature engineering technique for heart disease prediction with machine learning. IEEE Access, 11: 56214-56224. <https://doi.org/10.1109/ACCESS.2023.3281484>
- [7] Baumgartner, M., Veeranki, S.P.K., Hayn, D., Schreier, G. (2023). Introduction and comparison of novel decentral learning schemes with multiple data pools for privacy-preserving ECG classification. Journal of Healthcare Informatics Research, 7(3): 291-312. <https://doi.org/10.1007/s41666-023-00142-5>
- [8] Saikumar, K., Rajesh, V., Babu, B.S. (2022). Heart disease detection based on feature fusion technique with augmented classification using deep learning technology. Traitement du Signal, 39(1): 31-42. <https://doi.org/10.18280/ts.390104>
- [9] Yang, Z. (2022). Research on heart disease prediction method based on convolutional neural network. In 5th International Conference on Computer Information Science and Application Technology (CISAT 2022), pp. 1252-1261. <https://doi.org/10.1109/ACCESS.2023.3281484>
- [10] Tuli, S., Basumatary, N., Gill, S.S., et al. (2020). HealthFog: An ensemble deep learning-based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments. Future Generation Computer Systems, 104: 187-200. <https://doi.org/10.1016/j.future.2019.10.043>
- [11] Barandas, M., Famigliini, L., Campagner, A., et al. (2024). Evaluation of uncertainty quantification methods in multi-label classification: A case study with automatic diagnosis of electrocardiogram. Information Fusion, 101: 101978. <https://doi.org/10.1016/j.inffus.2023.101978>
- [12] Arooj, S., Rehman, S.U., Imran, A., Almuhaimeed, A., Alzahrani, A.K., Alzahrani, A. (2022). A deep convolutional neural network for the early detection of heart disease. Biomedicines, 10(11): 2796. <https://doi.org/10.3390/biomedicines10112796>
- [13] Almulihi, A., Saleh, H., Hussien, A.M., et al. (2022). Ensemble learning based on hybrid deep learning model for heart disease early prediction. Diagnostics, 12(12): 3215. <https://doi.org/10.3390/diagnostics12123215>
- [14] Modak, S., Abdel-Raheem, E., Rueda, L. (2022). Heart disease prediction using adaptive infinite feature selection and deep neural networks. In 2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC), Jeju Island, Korea, Republic of, pp. 235-240. <https://doi.org/10.1109/ICAIIIC54071.2022.9722652>
- [15] Banoth, R., Godishala, A.K., V.R., Yassin, H. (2022). A healthcare monitoring system for predicting heart disease

- through recurrent neural network. In 2022 IEEE 7th International Conference for Convergence in Technology (I2CT), Mumbai, India, pp. 1-7. <https://doi.org/10.1109/I2CT54291.2022.9824888>
- [16] Gjoreski, M., Gradisek, A., Budna, B., Gams, M., Poglajen, G. (2020). Machine learning and end-to-end deep learning for the detection of chronic heart failure from heart sounds. *IEEE Access*, 8: 20313-20324. <https://doi.org/10.1109/ACCESS.2020.2968900>
- [17] Vayadande, K., Golawar, R., Khairnar, S., et al. (2022). Heart disease prediction using machine learning and deep learning algorithms. In 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, pp. 393-401. <https://doi.org/10.1109/CISES54857.2022.9844406>
- [18] Chowdary, K.R., Bhargav, P., Nikhil, N., Varun, K., Jayanthi, D. (2022). Early heart disease prediction using ensemble learning techniques. *Journal of Physics: Conference Series*, 2325(1): 012051. <https://doi.org/10.1088/1742-6596/2325/1/012051>
- [19] Mir, H.Y., Singh, O. (2021). ECG denoising and feature extraction techniques – A review. *Journal of Medical Engineering & Technology*, 45(8): 672-684. <https://doi.org/10.1080/03091902.2021.1955032>
- [20] Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M.F., Ullah, N. (2022). A method for improving prediction of human heart disease using machine learning algorithms. *Mobile Information Systems*, 2022(1): 1-9. <https://doi.org/10.1155/2022/1410169>
- [21] Anitha, C., Rajkumar, S. (2023). An effective heart disease prediction method using extreme gradient boosting algorithm compared with convolutional neural networks. In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, pp. 2224-2228. <https://doi.org/10.1109/ICACCS57279.2023.10112952>
- [22] Yuan, X., Chen, J., Zhang, K., Wu, Y., Yang, T. (2022). A stable AI-based binary and multiple class heart disease prediction model for IoMT. *IEEE Transactions on Industrial Informatics*, 18(3): 2032-2040. <https://doi.org/10.1109/TII.2021.3098306>
- [23] Sarra, R.R., Dinar, A.M., Mohammed, M.A., Abdulkareem, K.H. (2022). Enhanced heart disease prediction based on machine learning and χ^2 statistical optimal feature selection model. *Designs*, 6(5): 87. <https://doi.org/10.3390/designs6050087>
- [24] Al Ahdal, A., Rakhra, M., Badotra, S., Fadhaeel, T. (2022). An integrated machine learning techniques for accurate heart disease prediction. In 2022 International Mobile and Embedded Technology Conference (MECON), Noida, India, pp. 594-598. <https://doi.org/10.1109/MECON53876.2022.9752342>
- [25] El-Hasnony, I.M., Elzeki, O.M., Alshehri, A., Salem, H. (2022). Multi-label active learning-based machine learning model for heart disease prediction. *Sensors*, 22(3): 1184. <https://doi.org/10.3390/s22031184>
- [26] Vasconcellos, M.E., Ferreira, B.G., Leandro, J.S., et al. (2023) Siamese convolutional neural network for heartbeat classification using limited 12-lead ECG datasets. *IEEE Access*, 11: 5365-5376. <https://doi.org/10.1109/ACCESS.2023.3236189>
- [27] Ghongade, O.S., Reddy, S.K.S., Tokala, S., Hajarathaiah, K., Enduri, M.K., Anamalamudi, S. (2023). A comparison of neural networks and machine learning methods for prediction of heart disease. In 2023 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur, India, pp. 1-7. <https://doi.org/10.1109/ICCT56969.2023.10076174>
- [28] Sree, S.N., Reddy, K.B.P., Mahanthi, B., Nandini, D.S.S. (2023). An analysis of heart disease prediction using machine learning. *International Journal for Multidisciplinary Research*, 5(3): 1-12. <https://doi.org/10.36948/ijfmr.2023.v05i03.3411>
- [29] Cocianu, C.L., Uscatu, C.R., Kofidis, K., Muraru, S., Văduva, A.G. (2023). Classical, evolutionary, and deep learning approaches of automated heart disease prediction: A case study. *Electronics*, 12(7): 1663. <https://doi.org/10.3390/electronics12071663>
- [30] Kachhawa, A., Hitt, J. (2022). An intelligent system for early prediction of cardiovascular disease using machine learning. *Journal of Student Research*, 11(3): 1-10. <https://doi.org/10.47611/jsrhs.v11i3.2989>
- [31] Khan, F., Yu, X., Yuan, Z., Rehman, A. (2023). ECG classification using 1-D convolutional deep residual neural network. *Plos One*, 18(4): e0284791. <https://doi.org/10.1371/journal.pone.0284791>
- [32] Varshini, G., Ramya, A., Sravya, C.L., Kumar, V., Shukla, B.K. (2023). Improving heart disease prediction of classifiers with data transformation using PCA and Relief feature selection. In 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, pp. 1644-1649. <https://doi.org/10.1109/ICEARS56392.2023.10085401>
- [33] Sen, K., Verma, B. (2023). Heart disease prediction using a soft voting ensemble of gradient boosting models, RandomForest, and Gaussian Naïve Bayes. In 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, pp. 1-7. <https://doi.org/10.1109/INCET57972.2023.10170399>
- [34] Khan, A., Qureshi, M., Daniyal, M., Tawiah, K. (2023). A novel study on machine learning algorithm-based cardiovascular disease prediction. *Health & Social Care in the Community*, 2023(1): 1406060. <https://doi.org/10.1155/2023/1406060>
- [35] Cheng, J., Zou, Q., Zhao, Y. (2021). ECG signal classification based on deep CNN and BiLSTM. *BMC Medical Informatics and Decision Making*, 21(1): 365. <https://doi.org/10.1186/s12911-021-01736-y>