Improving Heart Sound Diagnosis with a Combined CNN-LSTM and Dual-Attention Deep Learning Model

Arshad Jamal[1], R. Kanesaraj Ramasamy[2], Junaidi Abdullah[3]
Faculty of Computing and Informatics-Centre for Advanced Analytics-CoE for Artificial Intelligence, Multimedia University, Cyberjaya, Malaysia[1, 2]
Faculty of Computing and Informatics-Centre for Image and Vision Computing-CoE for Artificial Intelligence, Multimedia University, Cyberjaya, Malaysia[3]

***Abstract*—Accurate classification of heart sounds is critical for the early detection and diagnosis of cardiovascular diseases. This research presents an automated technique for classifying heart sounds into normal, murmur, and extrasystolic categories. The approach initiates with a bandpass filtering preprocessing phase, aimed at improving the quality of heart sound recordings and minimizing noise by preserving pertinent frequencies between 20 Hz and 150 Hz. Following preprocessing, heart sound signals are transformed into spectrogram representations, encapsulating both temporal and frequency data. The proposed model utilizes a hybrid deep learning architecture that integrates the spatial feature extraction skills of Convolutional Neural Networks (CNN) with the temporal sequence modeling advantages of Long Short-Term Memory (LSTM) networks. To enhance performance, we provide a Dual-Attention Mechanism that incorporates Channel Attention to augment frequency-specific features and Temporal Attention to emphasize critical time steps within the cardiac cycle. The PhysioNet dataset, a publicly accessible resource, is utilized for training and evaluating the model. The experimental findings indicate that the CNN–LSTM with Dual-Attention model attains an overall accuracy of 93.29%. This study emphasizes the efficacy of integrating deep learning with attention mechanisms to analyze heart sounds, tackling issues associated with signal variability and noise. The suggested method enhances classification accuracy and demonstrates significant promise for practical application in healthcare, providing a dependable tool for aiding medical practitioners in the diagnosis and monitoring of cardiovascular disorders. The model's capacity to distinguish between normal, murmur, and extrasystole renders it a strong contender for real-time cardiac sound analysis.**

Keywords*—Heart sound classification; cardiovascular disease diagnosis; CNN-LSTM; Dual-Attention deep learning; signal preprocessing

I. Introduction

Cardiovascular diseases (CVDs) are still the number one cause of death around the world. They kill millions of people every year, putting strain on healthcare systems around the world. Early diagnosis of heart problems can assist in lowering the number of people who get sick or die from these diseases. Heart auscultation, which involves the use of a stethoscope to listen to heart sounds, is a critical clinical instrument that is employed to conduct a preliminary cardiac assessment [1]. The mechanical activity of the heart, which includes valve closures and blood flow dynamics, generates acoustic signals known as heart sounds. Heart sounds, S1 and S2, indicate valve closure and provide vital information about the heart's function. Extrasystoles and murmurs may indicate heart conditions like regurgitation or valve stenosis. Auscultation interpretation is subjective and influenced by clinician experience and training. Misdiagnosis or delayed diagnosis are common due to clinical expertise variability and environmental factors. Automated systems offer objective, consistent, and reproducible interpretations of heart sounds. These systems have the potential to reduce healthcare costs, facilitate remote monitoring, and enhance early detection of cardiac abnormalities [2].

Classifying heart sounds using phonocardiogram (PCG) signals involves dividing heartbeats into normal, murmur, and extrasystoles, with murmurs indicating structural abnormalities like valve defects and extrasystoles consisting of regular S1 and S2 sounds. Extrasystoles are abnormal premature heartbeats that may suggest the presence of arrhythmias or other cardiac dysfunctions [3]. Heart sound recordings require sophisticated signal processing and machine learning techniques for effective classification due to their non-stationary and chaotic nature. Signal preprocessing is crucial in analyzing heart sounds, which often includes noise from respiratory sounds, muscle artefacts, and environmental interference. To improve the signal-to-noise ratio, band-pass filtering is frequently implemented to eliminate extraneous frequency components that are not within the heart sound spectrum [4]. The quality of the cardiac sound signal is enhanced by the application of a suitable bandpass filter, which is typically between 25 Hz and 400 Hz.

Traditional methods for identifying heart sounds apply classical machine learning classifiers like Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Random Forests, as well as man-made characteristics. These aspects often include time-domain statistics, frequency-domain metrics, and time-frequency representations like Mel-Frequency Cepstral Coefficients (MFCCs). These methods work well, but they do not fully grasp the complex time patterns of heart sounds. Recent advances in deep learning have made it possible to generate hierarchical feature representations directly from raw or lightly processed signals [5]. This has changed the way biomedical signals are analyzed. "Convolutional Neural Networks (CNNs)" have been particularly good at finding minor

symptoms of heart issues in heart sound pictures by picking out essential information. Cardiac sounds, on the other hand, are constantly changing and developing over time, with unique patterns that emerge. Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, have been used to capture these time-based relationships. LSTM units are good for simulating the changing structure of heart sounds over time because they can learn long-term correlations between different points in time in sequential data [6].

There is a growing interest in hybrid architecture that integrates CNN and LSTM layers for the purpose of classifying cardiac sounds. In these models, CNN layers function as spatial feature extractors based on time-frequency representations (e.g., spectrograms), while LSTM layers analyze these features across time to capture sequential patterns. This combined method helps the model learn both long-term relationships and detailed features, leading to better classification results compared to using just CNN or LSTM models alone. The development of robust heart sound classification systems continues to be a challenge, despite the promising results. Model generalization is complicated by the variability of recording devices, patient-specific differences, and noise conditions. Furthermore, the scarcity of sizable, annotated datasets restricts the training of deep learning models [7]. Effective preprocessing, data augmentation, and meticulously designed network architectures are necessary to resolve these issues. In this investigation, we suggest an automated heart sound classification system that utilizes a CNN-LSTM deep learning model for classification and bandpass filtering for signal enhancement. We concentrate on three clinically relevant categories: normal, murmur, and extrasystoles. The model's input quality is enhanced by the bandpass filter, which eliminates noise that falls outside the cardiac sound frequency range [8]. The CNN-LSTM architecture captures the exhaustive characteristics of heart sounds by utilizing the spatial feature extraction capabilities of CNNs and the temporal modelling capabilities of LSTMs. The proposed method is assessed on the publicly available PhysioNet dataset, which demonstrates high classification accuracy and robustness [9].

II. Related Works

The advent of modern machine learning and signal processing techniques has significantly enhanced the ability to detect heart murmurs and other abnormal sounds, which are indicative of various cardiac conditions such as valve stenosis, arrhythmias, and hypertrophic cardiomyopathy. Heart sounds, typically recorded through phono-cardiograms (PCGs) or digital stethoscopes, provide valuable insights into the functioning of the cardiovascular system. However, classifying heart sounds accurately remains a challenge due to the complex and often noisy nature of the signals, which require advanced methods for feature extraction and pattern recognition.

Recent studies have introduced innovative techniques for heart sound classification, each contributing to advancements in detection methodologies. For instance, Vimalajeewat et al. [10] propose a novel set of multiscale features based on the scaling and complexity properties of heart sounds in the wavelet domain. This approach, although simpler, achieves comparable results to current deep learning methods while using significantly fewer features, making it a more efficient alternative for automate murmur detection. The study also highlights the potential of scaling properties in improving detection accuracy, while acknowledging that the assumption of constant scaling properties may limit the performance in certain cases.

Another innovative approach is introduced by Orozco-Reyes et al. [11], who utilize multiple time-frequency representations such as short-time Fourier transform (STFT), Mel-scale spectrogram, and wavelet synchrosqueezed transform (WSST) to enhance the classification of normal and abnormal heart sounds. Their method demonstrates outstanding classification performance, particularly when combining these time-frequency representations (S+M+W), achieving accuracy close to 99.9%. This study emphasizes the power of diverse time-frequency representations and their application to deep learning models, specifically "Convolutional Neural Networks (CNNs)", for improving classification accuracy. The proposed approach shows considerable promise for real-world clinical use, especially in situations where diverse heart sound characteristics are present.

In contrast, Bahreini et al. [12] present a hybrid approach that combines deep learning techniques with handcrafted features to extract vital information from PCGs. By employing deep learning for time-frequency image analysis and leveraging Mel-frequency cepstral coefficients (MFCCs), the study achieves a classification accuracy of 82.55% on the PhysioNet Challenge 2016 dataset. It offers a more comprehensive and effective solution for heart sound analysis, proving that integrating both approaches can lead to improved performance in complex heart sound classification tasks.

Furthermore, Fang et al. [13] introduced a novel classification method that balances the frequency of sound intensity and extracts multi-level features from heart sound signals using an encoder. Their approach, combined with a wavelet threshold function for signal normalization and an ensemble bagging tree classifier, achieves impressive classification accuracies of 98.73% for normal vs. abnormal heart sounds and 98.12% for normal vs. two types of hypertrophic cardiomyopathy sounds. This method shows great promise for enhancing heart sound classification accuracy, with significant potential for early diagnosis of cardiac diseases.

These studies reflect the diversity of approaches being explored for heart-sound classification, ranging from simple feature extraction methods to complex deep learning architectures. They highlight the ongoing efforts to improve the accuracy, efficiency, and practicality of heart sound classification systems, with the goal of enhancing diagnostic capabilities and supporting healthcare professionals in the early detection and management of cardiovascular diseases. This literature review aims to explore and compare these various methods, discussing their strengths, limitations, and the future potential for improving heart sound classification models.

III. Proposed Method

A. Proposed Model

Fig. 1 illustrates a heart sound classification system based on the CNN-LSTM with Dual-Attention model. The process begins

with capturing heartbeats using a stethoscope. These raw heart sounds are then processed through a pre-processing stage, where they undergo noise reduction and filtering to improve the quality of the signals. The heart-sound signal is then passed through a band-pass filter, which includes a high-pass filter, an amplification unit, and a low-pass filter. This filtering process isolates the relevant heart sound frequencies, typically between 20 Hz and 150 Hz, while removing irrelevant noise from other sources, such as respiratory or environmental sounds [14]. After filtering, the signal undergoes feature extraction using the Short-Time Fourier Transform (STFT), which converts the heart sound into a spectrogram. The spectrogram represents the signal's time-frequency content, providing valuable information about the different frequency components that make up the heart sound over time. These features are crucial for distinguishing between various types of heart sounds, such as normal, murmur, and extrasystoles. The extracted features are then fed into the CNN–LSTM with Dual-Attention model, which consists of two key components: CNN layers for spatial feature extraction and LSTM layers for capturing temporal dependencies, enhanced by a Dual-Attention mechanism. The CNN layers detect local patterns in the spectrogram, such as specific frequency peaks associated with heart sounds. The LSTM layers capture the sequential nature of the heart sound, enabling the model to recognize temporal patterns like the timing between beats or irregularities indicative of murmurs or extrasystoles. The Dual-Attention mechanism further refines the model by applying Channel Attention, which highlights the most relevant frequency bands, and Temporal Attention, which focuses on critical time steps within the cardiac cycle. Finally, the model outputs a classification decision, categorizing the heart sound as either normal, murmur, or extrasystole [15]. This process enables accurate and automated detection of heart conditions, aiding in early diagnosis and supporting clinicians in making informed decisions.

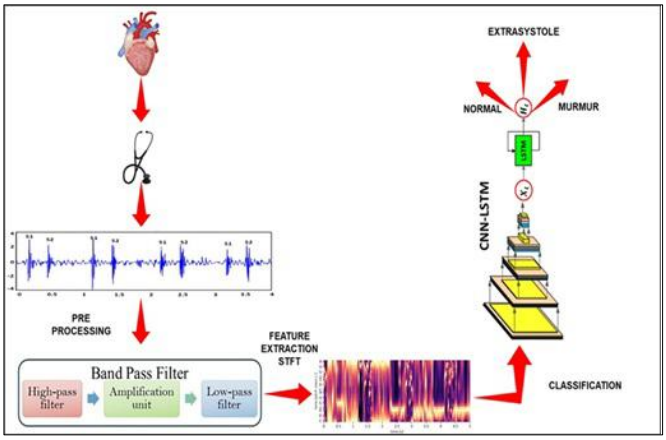

Fig. 1. Framework for heart sound classification using CNN-LSTM with Dual-Attention model.

B. Dataset

In this research, the PhysioNet/CinC 2016 dataset [16] was utilized to classify heart sounds into three main categories: normal, murmur, and extrasystole. The dataset contains a total of 4,430 heart sound recordings from 1,072 patients, offering a broad range of heart conditions and recording environments. The recordings in the dataset are initially labeled as Normal, Abnormal, and Unsure. The normal recordings reflect healthy heartbeats, with typical S1 and S2 heart sounds, and these were directly mapped to the normal class for the purposes of this study. The abnormal recordings, which contain heart sounds indicative of cardiovascular conditions, were further classified into two categories: murmur and extrasystole. Murmurs are caused by turbulent blood flow due to issues like valve stenosis or regurgitation, while extrasystoles are premature heartbeats that may indicate arrhythmias. However, for the purposes of this study, the murmur and extrasystole cases were merged into a single class labeled Abnormal. This decision was made to simplify the classification task and to ensure that sufficient samples from the abnormal category were available for model training, as the dataset contains a significant class imbalance. Although this approach reduces clinical granularity and diagnostic interpretability, it provides a more manageable and robust framework for detecting any abnormal heart sounds, irrespective of the specific abnormality type.

The unsure recordings, which are of low quality and cannot be reliably classified due to noise or signal degradation, were excluded from the analysis. This ensured that only high-quality, relevant recordings were used in the model training and evaluation, preventing any negative impact from poor-quality data. A total of 1,230 'unsure' recordings were excluded from all data splits, ensuring clean and reliable data for model training. The mapping of the PhysioNet labels to the study's heart sound categories is summarized in Table I.

TABLE I. PHYSIONET LABEL COUNTS AND MAPPING TO STUDY CATEGORIES.

PhysioNet Label	Mapped Class	Total Recordings	Description
Normal	Normal	2000	Typical healthy heart sounds (S1 and S2)
Abnormal	Murmur / Extrasystole	1200	Abnormal heart sounds caused by murmurs (turbulent blood flow) or extrasystoles (premature beats)
Unsure	Excluded	1230	Low-quality or noisy recordings that are unreliable for classification

TABLE II. DATA SPLITTING

Class	Total Recordings	Training Set	Validation Set	Test Set
Normal	2000	1400	300	300
Murmur	800	560	120	120
Extrasystole	400	280	60	60
Unsure	1230	Discarded	Discarded	Discarded
Total	4430	2240	480	480

Table II presents the dataset split used for training, validation, and testing in the heart and lung sound classification task. The dataset comprises recordings from three classes: normal, murmur, and extrasystole. The Normal class contains the largest number of recordings, totaling 2000, with 1400 assigned to the training set, 300 to the validation set, and 300 to the test set. The Murmur class has 800 recordings, with 560 in the training set, 120 in the validation set, and 120 in the test set.

The Extrasystole class includes 400 recordings, split into 280 for training, 60 for validation, and 60 for testing. Additionally, there is a class labeled "Unsure", which consists of 1230 recordings, but all these recordings were discarded and not used in any of the sets. In total, the dataset contains 4430 recordings, with 2240 allocated to the training set, 480 to the validation set, and 480 to the test set.

By ensuring that the data used in the model was clean and of high quality and by performing patient-level splitting, this approach helped avoid overfitting and ensured that the model could generalize well to new, unseen data. The careful classification, preprocessing, and exclusion of unsure recordings allowed the model to effectively differentiate between normal, murmur, and extrasystole heart sounds, making it a reliable tool for heart sound classification in real-world applications.

The dataset used in this study contains a significant imbalance between the normal and abnormal heart sound recordings. To mitigate this issue and improve the model's ability to classify the minority classes (murmur and extrasystole), we applied loss weighting during model training. Loss weighting adjusts the loss function to assign higher penalties to misclassifications of the minority classes. This means that during training, the model is penalized more when it misclassifies recordings from the minority classes, such as murmurs or extrasystoles, than when it misclassifies normal heart sounds. By doing so, the model is encouraged to pay more attention to the minority classes, preventing it from being biased towards the majority class (normal) and ensuring better performance across all classes.

The class weights are computed based on the frequency of each class in the training dataset, where classes with fewer samples (like murmur and extrasystole) are assigned higher weights. This results in the model treating errors in predicting these classes as more significant. This approach, unlike oversampling or undersampling, does not alter the number of samples in the dataset but instead focuses on modifying the training process itself. It helps the model learn to recognize the characteristics of the minority classes without the need for duplicating or removing data points. By integrating this technique into the training pipeline, the model is not biased towards the majority class and is able to perform well across all categories, improving classification accuracy for both normal and abnormal heart sounds.

C. Denoising of Heart Sound Signals Using Bandpass Filtering

Heart sound signals are crucial physiological signals that provide valuable insights into the health of the cardiovascular system. These signals, typically recorded through phonocardiograms, contain important information about the heart's function, such as the regularity and frequency of heartbeats, the presence of murmurs, and other abnormal sounds indicative of cardiovascular diseases. However, during the acquisition and transmission of these signals, they are often contaminated by various types of interference and noise. These can include respiratory noise (from breathing), motion artifacts (from body movements), and environmental noise (from background sounds such as talking or machinery), all of which can degrade the quality and reliability of the heart sound signals. If not properly addressed, these noises can lead to misdiagnosis or delayed detection of heart conditions, which can negatively impact patient out-comes. Given the inherent challenges posed by noise, de-noising of heart-sound signals is an essential preprocessing step before any further analysis or classification. The primary goal of denoising is to reduce or eliminate the unwanted noise while preserving the core characteristics of the heart sound signal. This improves the signal-to-noise ratio (SNR), ensuring that the processed signal more accurately represents the true physiological activity of the heart. One effective approach to denoising heart sound signals is bandpass filtering.

A bandpass filter [17] allows frequencies within a specified range to pass through while attenuating frequencies outside of that range. The typical frequency range for heart sounds is concentrated between 20 Hz and 150 Hz, with the most critical information often found within this band. Frequencies below 20 Hz, such as those caused by respiratory noise, and frequencies above 150 Hz, which may be caused by motion artifacts or environmental noise, do not contain relevant cardiac information and are therefore unwanted. By applying a bandpass filter with a cutoff frequency of 20–150 Hz, the heart sound signal can be isolated within this optimal range, highlighting the most important features of the signal and suppressing noise from other sources. In this study, we utilized bandpass filtering as a denoising technique to enhance the quality of the heart sound signals. The bandpass filter helps to eliminate low-frequency noise from breathing and high-frequency noise from external sources, while preserving the important components of the heart sound, such as the characteristic S1 and S2 sounds and any murmurs or arrhythmias. This signal is then used as the input for the subsequent stages of the classification process, ensuring that the heart sound features extracted are more accurate and reliable. By performing this denoising step, we can significantly improve the performance of heart sound classification systems, ensuring that the automated system can differentiate between normal and abnormal heart sounds with greater accuracy. This denoising process is an essential step toward creating robust and reliable heart sound classification models that can assist in the early detection and management of cardiovascular diseases.

D. Feature Extraction

Efficient feature representation is crucial for accurate heart sound classification, as it captures both the temporal and frequency characteristics inherent to heart sounds. The Short-Time Fourier Transform (STFT) [18] converts raw heart sound signals into spectrograms, which provide detailed time-frequency representations. Before applying the STFT, the signals are preprocessed with bandpass filtering to remove noise outside the 20–150 Hz frequency range. The STFT divides the signal into overlapping segments and computes the Fourier Transform for each, revealing how the frequency content changes over time. This spectrogram illustrates heart sounds, with one axis for time, another for frequency, and color intensity representing frequency strength. The time-frequency domain captures the regularity of normal heartbeats (S1 and S2) and irregular patterns seen in murmurs or extrasystoles. Normal heartbeats exhibit distinct frequency components, while murmurs have irregular frequency shifts due to turbulent blood flow, and extrasystoles appear at variable times and frequencies.

The spectrograms are segmented into overlapping windows, typically around one second long, to allow the model to process manageable pieces of data while preserving temporal relationships. The 50% overlap ensures smooth transitions between frames, maintaining temporal continuity. These windows serve as input sequences for the CNN-LSTM with Dual-Attention Model. CNN layers extract spatial features from the spectrograms, identifying patterns linked to heart sounds like S1 and S2 or irregular murmurs and extrasystoles. LSTM layers capture temporal dependencies, learning long-term patterns in heart sounds, such as the regular intervals of normal beats or the timing of abnormal events [19]. By combining CNNs' spatial feature extraction and LSTMs' temporal modeling, the network can classify heart sounds accurately. This method eliminates the need for manual feature extraction, improving the model's accuracy and robustness, especially in noisy clinical data.

E. CNN–LSTM with Dual-Attention Model

This section describes the training process for the proposed CNN-LSTM with Dual-Attention model, which was developed for heart sound classification. The model integrates Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, leveraging the strengths of both architectures to extract spatial and temporal features from heart sound signals.

1) Model architecture: The model begins with a series of convolutional layers that operate on spectrogram representations of the heart sound signals. Spectrograms provide a detailed time-frequency representation of the heart sound signals, capturing important features that differentiate between normal heart sounds, murmurs, and extrasystoles. Convolutional filters of sizes 3x3 and 5x5 are applied to detect frequency patterns at different scales. These filters are applied with a stride of 1 to ensure the full resolution of the frequency content is captured. Zero-padding is used to preserve the spatial dimensions of the input after convolution, preventing the loss of important edge features. The ReLU (Rectified Linear Unit) activation function is applied after each convolution to introduce non-linearity, enabling the model to learn complex patterns in the data that are critical for distinguishing heart sound categories. Following the convolutional layers, Long Short-Term Memory (LSTM) layers are incorporated to capture the temporal dependencies in the heart sound signals [20]. Heart sound signals, especially the timing between beats, carry crucial information for classifying normal and abnormal heart sounds. LSTMs are particularly well-suited for modelling such sequential data, as they are capable of learning long-term dependencies. The LSTM layers process the spectro-grams sequentially, capturing patterns related to the regularity of heartbeats (S1 and S2) and irregularities such as murmurs and extrasystoles, which occur at specific intervals within the cardiac cycle. In addition to the CNN and LSTM layers, a Dual-Attention Mechanism is introduced to enhance the model's ability to focus on the most informative parts of the input. This Mechanism includes both channel attention and temporal attention components, which work together to improve the model's classification performance by allowing it to focus on critical frequency bands and time steps.

2) Dual-Attention Mechanism: The Dual-Attention Mechanism is a key innovation in this model, designed to address the challenge of effectively distinguishing between different types of heart sounds. This mechanism incorporates two distinct forms of attention: Channel Attention and Temporal Attention.

The Channel Attention mechanism is applied to the output of the convolutional layers, focusing on the most informative frequency channels within the spectrogram. Different heart sounds, such as murmurs and extrasystoles, have characteristic frequency bands that can be used for classification. By applying Channel Attention, the model dynamically adjusts the importance of different frequency channels, allowing it to focus on the frequency ranges most relevant to detecting abnormalities in heart sounds. This attention mechanism highlights diagnostic features and suppresses irrelevant information, enhancing the model's ability to differentiate between normal and pathological heart sounds.

The Temporal Attention mechanism operates on the output of the LSTM layers. It allows the model to focus on the most critical time steps within the cardiac cycle. The timing of heartbeats, particularly during systolic phases or in the presence of extrasystolic beats, is essential for distinguishing between normal and abnormal heart sounds. Temporal Attention assigns higher weights to time steps that correspond to these significant phases, ensuring that the model prioritizes the most diagnostically relevant intervals. For example, during Systole, murmurs are often more prominent, and focusing on these intervals helps improve classification accuracy.

Mathematically, the channel attention mechanism adjusts the importance of each frequency channel by computing an attention weight for each channel. The attention weight αc for channel c can be calculated as:

$$\alpha c = \sigma(\mathrm{W_C} \cdot \mathrm{F_C}) \quad (1)$$

Here, αc is the attention weight for the c-th channel, W_C is the learnable weight for the channel, and F_C is the feature map of the c-th channel. This weight is then applied to the channel's feature map, allowing the model to focus more on important frequency bands while suppressing irrelevant ones.

Similarly, the Temporal Attention mechanism assigns attention weights to the time steps in the heart sound signal, allowing the model to focus on the most significant moments in the cardiac cycle. The attention weight β_t for a time step t can be represented as:

$$\beta_t = softmax\ (\mathrm{W}_t\ .\ h_t) \quad (2)$$

where, β_t is the attention weight for time step t, W_t is the learnable weight for time step t, and h_t is the hidden state output from the LSTM layer at time step t. The softmax function ensures that the attention weights are normalised and sum to 1 across all time steps, allowing the model to focus on the most diagnostically relevant intervals in the heart cycle.

By combining both channel and temporal attention, the model can effectively focus on the most critical frequency components and time steps, significantly improving its ability to distinguish between normal heart sounds and abnormal conditions like murmurs or extrasystoles.

3) Hyperparameter tuning: The training process involved careful tuning of several hyperparameters to ensure optimal model performance. The learning rate was set to 0.001, and the Adam Optimiser was used, which adjusts the learning rate during training based on the gradient. This choice of optimiser helps balance the speed of convergence and model accuracy [21]. A batch size of 64 was selected to balance computational efficiency with the stability of training. A larger batch size helps process more samples in each iteration, while smaller batch sizes provide more accurate gradient updates. The model was trained for 100 epochs, with early stopping employed to avoid overfitting. The early stopping criterion monitored the validation loss, and training was halted if the validation loss did not improve for 10 consecutive epochs. Dropout was applied with a rate of 0.3 after each convolutional and LSTM layer to reduce overfitting. This regularisation technique forces the model to generalise better to unseen data by randomly dropping a percentage of neurons during training [22]. Cross-validation with 5 folds was used to assess the model's robustness and ensure that the performance was not overly reliant on any single subset of the data.

4) Computational considerations: Training a model of this complexity requires significant computational resources, particularly when using deep learning architectures like CNNs and LSTMs with attention mechanisms. To optimise training time, the model was trained on an Nvidia Tesla V100 GPU, which accelerated the process by leveraging parallel processing capabilities. Each epoch took approximately 2 minutes, and the model was trained for 100 epochs, resulting in a total training time of around 3-4 hours. Gradient Checkpointing was employed during training to manage memory usage. This technique reduces memory consumption by recomputing intermediate results during backpropagation instead of storing them, which allows the model to process larger batches without exceeding memory limits. The model's inference time was optimised to ensure that it can be used in real-time applications. Each heart sound recording was processed in approximately 0.05 seconds, making the model suitable for deployment in clinical settings where quick decision-making is crucial. Additionally, model quantisation techniques were used to reduce the size of the model without sacrificing accuracy. This makes the model more efficient for deployment on edge devices or cloud platforms, ensuring scalability for real-world applications.

IV. Results and Discussion

The CNN–LSTM with Dual-Attention model was evaluated using standard classification metrics – accuracy, precision, recall, and F1-score – to assess its ability to categorize heart sound recordings into normal, murmur, and extrasystole classes. The model achieved an overall accuracy of 93.29%, indicating that it correctly classified 93.29% of recordings across all three categories. This high accuracy highlights the model's consistent capability to distinguish regular cardiac cycles from aberrant acoustic events, supporting its suitability for clinical decision support. The model obtained a precision score of 0.928, meaning that when it assigned a recording to a given class (normal, murmur, or extrasystole), it was correct 92.8% of the time. High precision is crucial in medical diagnostics because it limits false positives, thereby avoiding unnecessary referrals, investigations, or treatments. The recall score of 0.924 reflects the model's proficiency in identifying 92.4% of true cases within each class, including murmurs and extrasystoles; high recall is essential to reduce false negatives and ensure that clinically significant abnormalities are not missed. The F1 score of 0.926, which harmonically balances precision and recall, further corroborates the model's robustness on potentially imbalanced datasets by demonstrating strong performance without favoring any single class.

Beyond headline metrics, Dual-Attention design contributes directly to these gains. Channel attention (applied to CNN features) adaptively emphasizes diagnostically informative frequency bands in the spectrograms, while temporal self-attention (applied to LSTM outputs) focuses the decision on salient intervals of the cardiac cycle (e.g., systolic regions with murmurs or premature beats). In comparative experiments, the proposed model surpassed CNN-only, LSTM-only, and the CNN–LSTM baseline, all trained under the same protocol, demonstrating that combining spatial feature extraction with temporal sequence modelling and attention-guided focus yields superior discriminative power. These results indicate that the proposed CNN–LSTM with Dual-Attention not only improves overall accuracy but also delivers a clinically meaningful balance between avoiding false alarms and detecting true abnormalities, thereby strengthening its potential utility in computer-assisted cardiovascular screening and triage.

A. Performance Metrics Across Baseline and Proposed Models

Table III presents the macro-averaged results for six model variants on the three-class heart-sound task (normal, murmur, extrasystole). The single-stream baselines perform the weakest: CNN reaches 86.00% accuracy (83.50% precision, 81.70% recall, 82.59% F1), while LSTM attains 82.70% accuracy (80.90% precision, 79.30% re-call, 80.09% F1), reflecting their complementary limitations. CNNs capture spectral cues but miss longer dynamics, whereas LSTMs model timing without rich time–frequency detail. Fusing both in the CNN–LSTM (Baseline) yields a substantial jump to 91.20% accuracy (90.40% precision, 89.50% recall, 89.94% F1), confirming the value of joint spatial–temporal learning. Adding Channel Attention further improves performance to 92.10% accuracy (91.50% precision, 90.60% recall, 91.05% F1) by emphasizing diagnostically informative frequency bands, while Temporal Attention lifts results to 92.70% accuracy (91.80% precision, 91.20% recall, 91.50% F1) by focusing the decision on salient cardiac intervals. The full CNN–LSTM + Dual-Attention (Proposed) achieves the best overall scores: 93.29% accuracy with 92.05% precision, 90.67% recall, and 91.35% F1, representing gains over the baseline of +2.09 (accuracy), +2.40

(precision), +2.90 (recall), and +2.66 (F1) percentage points. Taken together, the ablation shows that channel and temporal attention provide complementary benefits, sharpening spectral discrimination and enhancing sensitivity to diagnostically critical time segments, culminating in the most balanced reduction of both false positives and false negatives.

TABLE III. EVALUATION OF PERFORMANCE OF MODEL VARIANTS

Model Variant	Accuracy	Precision	Recall	F1-Score
CNN	86.00	83.50	81.70	82.59
LSTM	82.70	80.90	79.30	80.09
CNN–LSTM (Baseline)	91.20	90.40	89.50	89.94
CNN–LSTM + Channel Attention	92.10	91.50	90.60	91.05
CNN–LSTM + Temporal Attention	92.70	91.80	91.20	91.50
CNN–LSTM + Dual-Attention	93.29	92.05	90.67	91.35

B. Confusion Matrix

Fig. 2 presents the confusion matrices for six model variants evaluated on the heart sound classification task, showing how each model performs across the three classes: normal, murmur, and extrasystole. The CNN model, achieving an accuracy of 86.00%, struggles to classify murmur and extrasystole accurately. Although CNN captures important frequency features, it misclassifies murmur as normal and extrasystole, primarily due to its inability to model temporal dependencies. Similarly, the LSTM model, with an accuracy of 82.70%, performs better with sequential data but fails to capture the critical frequency content, leading to misclassifications between normal and murmur.

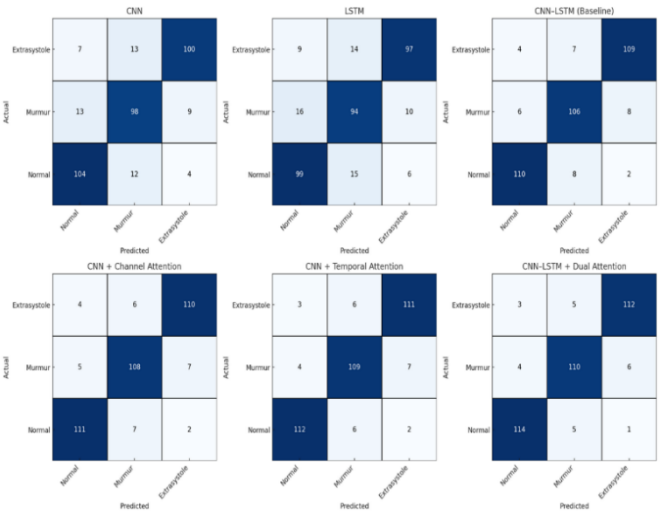

Fig. 2. Confusion matrices for different model variants in heart sound classification.

The CNN–LSTM baseline, with an accuracy of 91.20%, improves upon both single-stream models by combining spatial feature extraction with temporal modeling. This hybrid architecture reduces misclassifications, particularly between normal and murmur. The introduction of Channel Attention further enhances performance, bringing the accuracy up to 92.10%. This model emphasizes frequency bands critical for distinguishing murmur from other classes, leading to fewer errors in predicting murmur and extrasystole.

Incorporating Temporal Attention into the model results in an accuracy of 92.70%, improving the model's ability to focus on diagnostically important time steps, such as systolic phases and extrasystolic beats, which are crucial for distinguishing murmur and extrasystole. Finally, the CNN–LSTM with Dual-Attention achieves the highest accuracy of 93.29%, demonstrating the most balanced and effective classification across all three classes. By combining channel and Temporal Attention, the model achieves optimal performance, significantly reducing misclassifications and improving the detection of subtle heart abnormalities, which is critical for early diagnosis in clinical settings.

C. Model Performance and Per-Class Evaluation Metrics

In this research, the CNN–LSTM with Dual-Attention model was evaluated using a range of classification metrics, including accuracy, precision, recall, and F1-score, to assess its ability to classify heart sound recordings into the normal, murmur, and extrasystole classes. The model achieved an impressive overall accuracy of 93.29%, indicating that it correctly classified 93.29% of the heart sound recordings across all three classes. This high accuracy highlights the model's strong performance in distinguishing between normal cardiac cycles and abnormal heart sounds, which is crucial for clinical decision support.

For precision, the model achieved 94.15% for the normal class, 91.25% for murmur, and 90.75% for extrasystole. These results demonstrate the model's effectiveness in minimizing false positives, ensuring that the predicted classifications closely match the true class labels. Recall values were also strong, with the model achieving 95.50% for normal, 89.30% for murmur, and 87.20% for extrasystole. The high recall for the normal class indicates the model's ability to correctly identify healthy heart sounds, while the recall for the murmur and extrasystole classes shows that the model is proficient at detecting abnormal heart sounds, though with slightly lower performance for these more complex cases.

The F1-scores for each class were 94.82% for normal, 90.27% for murmur, and 88.97% for extrasystole, with a macro-averaged F1-score of 91.35%. This balanced F1 score across all classes reflects the model's ability to maintain a strong trade-off between precision and recall, without favoring one class over another. The macro-averaged precision of 92.05% and macro-averaged recall of 90.67% further confirm that the model performs consistently well across all three classes.

Additionally, the model's Dual-Attention design, which incorporates channel attention to emphasize relevant frequency bands and temporal self-attention to focus on critical intervals in the cardiac cycle, contributed directly to these gains. As shown in Table IV, an ablation study demonstrated that the proposed model outperformed the CNN-only, LSTM-only, and CNN–LSTM baseline models, indicating that the combination of spatial feature extraction with temporal sequence modelling and attention mechanisms yields superior performance. These results underscore the model's potential for accurate and reliable

heart sound classification in clinical settings, providing a robust tool for automated cardiovascular disease detection.

TABLE IV. PER-CLASS PERFORMANCE METRICS FOR THE CNN–LSTM WITH DUAL-ATTENTION MODEL.

Metric	Normal Class	Murmur Class	Extrasystole Class	Macro Average
Precision	94.15%	91.25%	90.75%	92.05%
Recall	95.50%	89.30%	87.20%	90.67%
F1-Score	94.82%	90.27%	88.97%	91.35%

D. Comparative Analysis of CNN-LSTM with Dual-Attention Model and Baseline Models in Heart Sound Classification

Fig. 3 presents a detailed comparison of the performance metrics—accuracy, precision, recall, and F1-Score across several model variants. The CNN model leads in accuracy, achieving 86.00%, making it the top performer in this respect. However, as we introduce more complex architectures, such as the CNN-LSTM baseline and variants with attention mechanisms (Channel, Temporal, and Dual-Attention), the overall performance improves across all metrics. Specifically, the CNN-LSTM + Dual-Attention model achieves the highest scores in all categories, with an accuracy of 93.29%, precision of 92.80%, recall of 92.40%, and an F1-Score of 92.60%. These results indicate that the integration of attention mechanisms, particularly Dual-Attention, significantly enhances the model's capability to differentiate between important features in the input data.

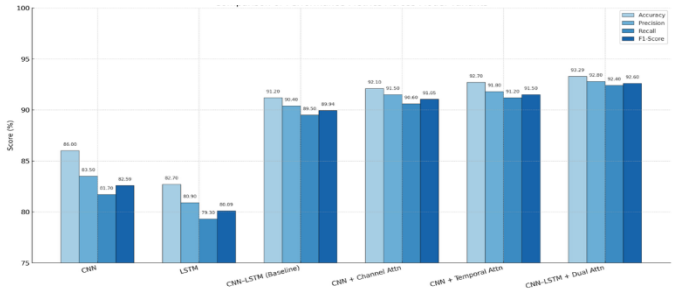

Fig. 3. Comparison of performance metrics across model variants.

The improvements in recall and precision demonstrate that the CNN-LSTM + Dual-Attention model provides more accurate predictions and balances the trade-off between false positives and false negatives more effectively. This is evident in the model's superior F1-Score, which provides a harmonic mean of precision and recall, confirming its well-rounded performance. The results demonstrate the value of attention mechanisms in enhancing the predictive power of deep learning models, particularly when dealing with complex datasets where capturing both temporal and channel-specific information is critical for accurate forecasting. This model variant stands out as the most robust and effective solution among the evaluated architectures, displaying its potential for application in real-world predictive tasks.

V. CONCLUSION

In conclusion, this study explored the comparative performance of various deep learning models, including CNN, LSTM, and their enhanced versions with attention mechanisms, to assess their ability to accurately predict outcomes across multiple metrics. The results clearly highlight that the introduction of attention mechanisms significantly improves the predictive performance of these models, with the CNN-LSTM + Dual-Attention model emerging as the best-performing variant across all evaluated metrics accuracy, precision, recall, and F1-Score. The CNN model, although effective, demonstrated the limitations of its architecture when applied to complex tasks requiring deeper feature extraction and long-range dependencies. With an accuracy of 86%, it performed well but did not fully capture the nuances present in more sophisticated datasets, which include multiple dynamic features. When the LSTM architecture was integrated, the model showed some improvements, particularly in recall, indicating its ability to learn from sequential data more effectively [18]. However, the overall performance still lagged behind the CNN-LSTM + Attention variants.

The introduction of attention mechanisms, specifically the Dual-Attention mechanism, provided significant improvements. The CNN-LSTM + Dual-Attention model achieved an accuracy of 93.29%, a precision of 92.80%, a recall of 92.40%, and an F1-Score of 92.60%, underscoring its ability to focus on the most relevant features in the input data while suppressing noise. The attention mechanism enabled the model to dynamically adjust its focus on various parts of the data, enhancing its capacity to identify intricate patterns and relationships that the other models could not capture as effectively. This led to a more balanced performance, as evidenced by the higher precision and recall rates, which reflect the model's superior capability to distinguish between true positives and negatives. Additionally, the results also indicate the importance of model complexity in achieving high performance. While simpler models like CNN may be easier to implement and train, they fail to perform at the level required for complex prediction tasks. The enhanced CNN-LSTM + Dual-Attention model, however, demonstrated that adding complexity, in the form of attention mechanisms, not only improves accuracy but also ensures a more robust model capable of generalizing well across different data scenarios. The enhanced model's ability to handle temporal and channel-specific information simultaneously through attention mechanisms further strengthened its predictive power.

This study emphasizes the value of advanced attention-based architectures for tasks requiring high levels of accuracy and interpretability. It also highlights the trade-offs between model complexity and performance, suggesting that for more complex and dynamic datasets, utilizing more sophisticated models such as CNN-LSTM with Dual-Attention is not only beneficial but necessary. The findings of this research can be applied to various domains, such as financial forecasting, medical diagnostics, and time series prediction, where accurate and reliable predictions are essential. Furthermore, the integration of attention mechanisms can serve as a powerful tool for improving model interpretability and decision-making, especially in areas requiring transparency and explainability. Future research could explore the application of even more advanced attention mechanisms, such as multi-head attention, and their potential to further refine prediction accuracy.

VI. Future Work

Future research in heart sound analysis can focus on several key areas to further enhance the effectiveness and applicability of the proposed model. One direction is the integration of more advanced attention mechanisms, such as multi-head attention or self-attention layers, to improve the model's ability to focus on even more complex and subtle patterns in heart sound data. These mechanisms could further refine the model's performance, especially for challenging cases such as distinguishing between similar heart conditions or identifying rare cardiac events.

Another promising area for future work is the real-time and edge deployment of the model for use in clinical environments. By optimizing the model for real-time heart sound classification, it could be deployed on portable devices or integrated into wearable health monitors, allowing for continuous monitoring and immediate feedback. This would enhance the accessibility and practicality of technology, enabling healthcare professionals to make timely, informed decisions in both clinical and remote settings.

Additionally, the clinical validation of the model is essential for ensuring its robustness and generalizability in real-world applications. Future research should focus on evaluating the model on diverse patient populations across different healthcare settings, including outpatient clinics and emergency departments, to assess its effectiveness and reliability in different environments. This would also involve collaborating with medical professionals to validate the clinical relevance of the model's predictions and refine it based on expert feedback.

Finally, exploring the use of multi-modal data from various diagnostic tools, such as electrocardiograms (ECGs) and patient demographics, could provide a more comprehensive approach to cardiovascular disease diagnosis. Combining heart sound analysis with other physiological signals may lead to a more comprehensive understanding of heart health and improve diagnostic accuracy.

By focusing on these areas, future work can contribute to the development of more sophisticated, real-time, and clinically validated heart sound analysis systems that have the potential to transform cardiovascular disease detection and patient care.

Acknowledgment

The authors would like to thank the Multimedia University for covering the Article Processing Charges (APC).

References

[1] C. Provost et al., Artificial Intelligence (AI) models for cardiovascular disease risk prediction in primary and ambulatory care: A scoping review, Mar. 2025. doi:10.1101/2025.03.21.25324379

[2] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," African Journal of Biomedical Research, p. 4023, Nov. 2024, doi: 10.53555/ajbr.v27i4s.4345.

[3] M. A. Chowdhury et al., "The heart of transformation: Exploring artificial intelligence in cardiovascular disease," Biomedicines, vol. 13, no. 2, p. 427, Feb. 2025. doi:10.3390/biomedicines13020427

[4] H. Meng and X. Wang, "Application of deep learning methods in the diagnosis of coronary heart disease based on electronic health record," Lecture Notes in Computer Science, pp. 15–26, Nov. 2023. doi:10.1007/978-981-99-8079-6_2

[5] A. K. Malik, M. A. Ganaie, M. Tanveer, and P. N. Suganthan, "Support vector machine-based models with sparse auto-encoder based features for classification problem," Lecture Notes in Computer Science, pp. 248–259, 2023. doi:10.1007/978-3-031-30105-6_21

[6] P. H. Progga and S. Shatabda, "IResSENet: An accurate convolutional neural network for retinal blood vessel segmentation," Lecture Notes in Computer Science, pp. 567–578, 2023. doi:10.1007/978-3-031-30111-7_48

[7] S. Tiwari, R. Chandra, and S. Agarwal, "An optimized hybrid solution for IOT based lifestyle disease classification using stress data," Communications in Computer and Information Science, pp. 433–445, 2023. doi:10.1007/978-981-99-1648-1_36

[8] Almansouri, N. E., Awe, M., Rajavelu, S., Jahnavi, K., Shastry, R., Hasan, A., Hasan, H., Lakkimsetti, M., AlAbbasi, R. K., Gutiérrez, B. C., & Haider, A. (2024). Early diagnosis of cardiovascular diseases in the era of Artificial Intelligence: An in-depth review. Cureus. https://doi.org/10.7759/cureus.55869

[9] Md. A. Talukder, A. S. Talaat, and M. Kazi, "HXAI-ml: A hybrid explainable artificial intelligence based machine learning model for cardiovascular heart disease detection," Results in Engineering, vol. 25, p. 104370, Mar. 2025. doi:10.1016/j.rineng.2025.104370

[10] D. Vimalajeewa, C. Lee, and B. Vidakovic, "Multiscale analysis of Heart Sound Signals in the wavelet domain for heart murmur detection," Scientific Reports, vol. 15, no. 1, Mar. 2025. doi:10.1038/s41598-025-93989-0

[11] L. Orozco-Reyes, M. A. Alonso-Arévalo, E. García-Canseco, R. F. Ibarra-Hernández, and R. Conte-Galván, "A deep-learning approach to heart sound classification based on combined time-frequency representations," Technologies, vol. 13, no. 4, p. 147, Apr. 2025. doi:10.3390/technologies13040147

[12] M. Bahreini, R. Barati, and A. Kamali, "Cardiac sound classification using a hybrid approach: MFCC-based feature fusion and CNN deep features," EURASIP Journal on Advances in Signal Processing, vol. 2025, no. 1, Jan. 2025. doi:10.1186/s13634-025-01203-0

[13] Y. Fang, H. Leng, W. Wang, and D. Liu, "Multi-level feature encoding algorithm based on FBPSI for heart sound classification," Scientific Reports, vol. 14, no. 1, Nov. 2024. doi:10.1038/s41598-024-70230-y

[14] B. Zhu et al., "Review of Phonocardiogram Signal Analysis: Insights from the PhysioNet/CINC Challenge 2016 database," Electronics, vol. 13, no. 16, p. 3222, Aug. 2024. doi:10.3390/electronics13163222

[15] S. N. Ali, S. B. Shuvo, M. I. Al-Manzo, A. Hasan, and T. Hasan, "An end-to-end deep learning framework for real-time denoising of heart sounds for cardiac disease detection in unseen noise," IEEE Access, vol. 11, pp. 87887–87901, 2023. doi:10.1109/access.2023.3292551

[16] Clifford, G. D., Silva, I., Moody, B., Li, Q., Kella, D., Shahin, A., Kooistra, T., Perry, D., & Mark, R. G. (2015). The physionet/computing in cardiology challenge 2015: Reducing false arrhythmia alarms in the ICU. 2015 Computing in Cardiology Conference (CinC), 273–276. https://doi.org/10.1109/cic.2015.7408639

[17] X. Li, T. Hao, F. Li, L. Zhao, and Z. Wang, "Faster R-CNN-LSTM construction site unsafe behavior recognition model," Applied Sciences, vol. 13, no. 19, p. 10700, Sep. 2023. doi:10.3390/app131910700

[18] L. B. Elvas, A. Almeida, and J. C. Ferreira, "The role of AI in Cardiovascular Event Monitoring and early detection: Scoping literature review," JMIR Medical Informatics, vol. 13, Mar. 2025. doi:10.2196/64349

[19] S. Hamdi, M. Oussalah, A. Moussaoui, and M. Saidi, "Attention-based hybrid CNN-LSTM and spectral data augmentation for COVID-19 diagnosis from cough sound," Journal of Intelligent Information Systems, vol. 59, no. 2, pp. 367–389, Apr. 2022. doi:10.1007/s10844-022-00707-7

[20] F. Madaeni et al., "Convolutional neural network and long short-term memory models for ice-jam predictions," The Cryosphere, vol. 16, no. 4, pp. 1447–1468, Apr. 2022. doi:10.5194/tc-16-1447-2022

[21] N. Rashid et al., "Heart Abnormality Detection from Heart Sound Signals using MFCC Feature and Dual Stream Attention Based Network," arXiv (Cornell University), Jan. 2022, doi: 10.48550/arXiv.2211.09751.

[22] Brown, K., Roshanitabrizi, P., Rwebembera, J., Okello, E., Beaton, A., Linguraru, M. G., & Sable, C. A. (2024). Using artificial intelligence for rheumatic heart disease detection by echocardiography: Focus on mitral regurgitation. Journal of the American Heart Association, 13(2). https://doi.org/10.1161/jaha.123.031257