Deep Learning in Heart Murmur Detection: Analyzing the Potential of FCNN vs. Traditional Machine Learning Models

Hajer Sayed Hussein[1], Hussein AlBazar[2]*, Roxane Elias Mallouhy[3]*, Fatima Al-Hebshi[4]
Faculty of Computer Studies, Arab Open University, Saudi Arabia[1,2,4]
Al Yamama University, Saudi Arabia[3]

***Abstract*—This research investigates the performance of machine learning and deep learning models in detecting heart murmurs from audio recordings. Using the PhysioNet Challenge 2016 dataset, we compare several traditional machine learning models—Support Vector Machine, Random Forest, AdaBoost, and Decision Tree—with a Fully Convolutional Neural Network (FCNN). The findings indicate that while traditional models achieved accuracies between 0.85 and 0.89, they faced challenges with data complexity and maintaining a balance between precision and recall. Ensemble methods such as Random Forest and AdaBoost demonstrated improved robustness but were still outperformed by deep learning approaches. The FCNN model, leveraging artificial intelligence, significantly outperformed all other models, achieving an accuracy of 0.99 with a precision of 0.94 and a recall of 0.96. These results highlight the potential of AI-driven cardiovascular diagnostics, as deep learning models exhibit superior capability in identifying intricate patterns in heart sound data. Our findings suggest that deep learning models offer substantial advantages in medical diagnostics, particularly for cardiovascular diagnostics, by providing scalable and highly accurate tools for heart murmur detection. Future work should focus on improving model interpretability and expanding dataset diversity to facilitate broader adoption in clinical settings.**

Keywords*—*Heart murmur detection; machine learning; deep learning; cardiovascular diagnostics; artificial intelligence; physioNet dataset

I. INTRODUCTION

Artificial Intelligence (AI) has profoundly impacted the medical field, particularly in its ability to rapidly process and analyze vast datasets with unprecedented accuracy and speed. This capability has revolutionized healthcare by enabling earlier and more precise diagnoses, the development of personalized treatment plans, and overall improved patient outcomes. Among the most promising applications of AI is its role in cardiac care, specifically in the detection and analysis of heart murmurs—abnormal heart sounds that can be indicative of various cardiovascular conditions. Heart murmurs serve as critical indicators of underlying heart issues, ranging from benign anomalies to severe, life-threatening diseases.

The heart produces characteristic "lub-dub" sounds as its valves close during the pumping of blood. A heart murmur is an additional sound detected during this process, often signaling potential turbulence in blood flow. While many murmurs are benign, others can signal structural abnormalities such as malfunctioning valves or congenital defects. Precise and early detection is crucial to facilitate timely intervention, significantly improving patient prognosis and reducing mortality rates. Cardiovascular diseases (CVD), many related to heart murmurs, are the leading cause of death globally, accounting for approximately 19.91 million deaths in 2021 [1]. Moreover, CVD represented 12% of total U.S. health expenditures from 2019 to 2020, making it the most costly diagnostic group [2]. A substantial portion of these fatalities could potentially be prevented with earlier detection and treatment. Heart murmurs are also a leading cause for referral to pediatric cardiologists, with studies indicating that up to 72% of children will experience a murmur at some point during their childhood or adolescence [3]. While some murmurs resolve over time, others may persist into adulthood, requiring ongoing evaluation and management.

Traditional methods of detecting heart murmurs, such as physical examination with a stethoscope, have several limitations compared to AI-based detection techniques. The accuracy of traditional auscultation heavily depends on the clinician's experience, leading to potential human error and variability in interpretation. This variability can result in inconsistent diagnoses and treatment plans. Faint or position-specific murmurs may be missed, and traditional methods do not provide quantitative data about the murmur's characteristics, limiting the ability to track changes over time. Moreover, traditional detection relies on the physical presence of both the patient and healthcare provider, making it less adaptable to remote monitoring. Complex heart conditions with subtle or mixed murmurs can be particularly challenging to diagnose accurately with a stethoscope alone. Additionally, auscultation can be time-consuming, potentially leading to rushed assessments in busy clinical settings. Furthermore, traditional methods do not easily integrate with other patient data, such as echocardiograms or electronic health records, whereas AI-based systems can combine multiple data sources for a more comprehensive diagnosis.

Despite advancements in Artificial Intelligence (AI) for medical diagnostics, there remains a gap in evaluating the performance of these models for heart murmur detection. While some studies have explored traditional ML approaches, there is limited research comparing these methods to deep learning architectures, such as Fully Convolutional Neural Networks (FCNN), in the specific context of heart murmur detection. This study aims to bridge this gap by systematically comparing the effectiveness of traditional ML models—Support Vector Machine, Random Forest, AdaBoost, and Decision Tree—against an FCNN model using heart sound recordings from the PhysioNet Challenge 2016 dataset. The objective is to determine whether deep learning provides significant improvements over traditional ML techniques in detecting heart murmurs and enhancing diagnostic accuracy.

Hence, integrating AI in heart murmur detection offers a transformative solution to these limitations by providing a more consistent and objective analysis. AI-powered diagnostic tools can detect subtle patterns in heart sound recordings that may not be discernible to the human ear, thereby enhancing diagnostic precision and ensuring that at-risk patients are identified earlier. AI's ability to reduce human error, process large volumes of data, and provide real-time diagnostic support across diverse healthcare settings offers significant advantages over traditional approaches. Despite these benefits, traditional methods continue to dominate clinical practice due to their reliance on the expertise and judgment of physicians, which can lead to variability in diagnosis and patient outcomes. The accuracy of traditional diagnostics often depends on the clinician's experience, and the processes can be time-intensive, potentially lacking the precision necessary for early detection.

To address these challenges, this study aims to develop an automated system for heart murmur detection utilizing machine learning (ML) and deep learning (DL) techniques. Leveraging a dataset from the PhysioNet Challenge 2016, the research will focus on the extraction of relevant features from heart sound recordings using Mel-Frequency Cepstral Coefficients (MFCC). MFCC is selected for its proven efficacy in capturing the essential characteristics of audio signals, making it an optimal choice for heart sound analysis. Following feature extraction, various ML and DL models will be implemented and trained on the dataset, with the goal of evaluating their accuracy and effectiveness in detecting heart murmurs. Moreover, this research seeks to demonstrate the significant potential of AI in enhancing the early detection of heart murmurs, ultimately leading to improved patient outcomes and a reduction in the global burden of cardiovascular diseases. By advancing the development of these sophisticated models, the study aims to contribute to the creation of a more reliable, accurate, and accessible diagnostic tool for heart murmur detection, thereby improving healthcare delivery and patient care.

The remainder of this article is organized as follows: Firstly, Section II provides a comprehensive review of the state-of-the-art methods in heart murmur detection, particularly highlighting the advancements and challenges associated with applying machine learning and deep learning techniques. Subsequently, Section III delves into the methodology employed in this study, where the processes of data collection, preprocessing, feature extraction, and the application of various ML and DL models are thoroughly detailed. Following this, Section IV presents the results, offering a comparative analysis of the models' performances while discussing their implications for clinical practice. Lastly, Section V concludes the article with a summary of the key findings, accompanied by an exploration of limitations and recommendations for future research directions.

II. State-of-the-Art

The advancements in AI for heart murmur detection and diagnosis have led to a diverse range of studies employing various ML and DL techniques to improve the accuracy and reliability of these methods. The Multi-Kernel Residual Convolutional Neural Network (MK-RCNN) model stands out as a significant innovation, capturing multi-scale features through multi-kernel convolutional networks and utilizing residual learning for deeper feature extraction. This model achieved an impressive 98.33% accuracy on three datasets, making it a promising tool for reliable heart murmur diagnosis in primary healthcare settings [4]. Complementing this approach, a comprehensive review on machine learning-based analysis of PCG signals underscores the importance of feature extraction and data quality in enhancing diagnostic accuracy. This review explores how supervised, unsupervised, and deep learning techniques have been effectively applied to heart sound analysis, significantly improving the accuracy of cardiovascular disease diagnosis [5].

In parallel, the development of novel real-time detection methods, such as FunnelNet, has demonstrated the efficiency of combining traditional and depthwise separable convolutional networks for heart murmur detection. FunnelNet employs continuous wavelet transform (CWT) for feature extraction and integrates SqueezeNet, a Bottleneck layer, and ExpansionNet, achieving state-of-the-art performance with 99.70% accuracy on four public datasets [6]. This method's suitability for resource-constrained devices highlights its potential for accessible medical services. Additionally, the exploration of general-purpose audio representations pre-trained on large-scale datasets, like the Masked Modeling Duo (M2D), has shown the effectiveness of self-supervised learning methods in heart sound analysis, with ensembling techniques further improving diagnostic outcomes [7].

Building on these advancements, other studies have focused on the classification of heart murmur quality. A study employing deep neural networks to classify heart murmur quality as harsh or blowing utilized a CNN with channel attention and GRU networks to extract features from log-Mel spectrograms, followed by a Feature Attention module to weight features across segments [8]. This model achieved 73.6% accuracy, with F1-scores of 76.8% for harsh murmurs and 67.8% for blowing murmurs, illustrating the nuanced capabilities of AI in analyzing heart sound characteristics. Moreover, traditional heart sound classification methods, which often depend on ECG-labeled PCGs or feature extraction from mel-scale frequency cepstral coefficients (MFCC), have seen significant improvements with the introduction of capsule neural networks (CapsNet) [9]. CapsNet enhances feature representation through iterative dynamic routing, achieving validation accuracies of 90.29% and 91.67%, thus offering a robust alternative for heart murmur detection.

Further innovation in this field includes the development of CardioXNet, a lightweight CRNN architecture designed to detect five cardiac conditions using raw PCG signals. CardioXNet combines representation learning with parallel CNN pathways for feature extraction and sequence residual learning using bidirectional LSTMs, capturing temporal features with high accuracy and low computational requirements [10]. This model's applicability in low-resource settings on mobile devices is particularly noteworthy. Alongside these advancements, a study exploring general-purpose audio representations pre-trained on large-scale datasets for heart murmur detection introduced the self-supervised learning method Masked Modeling Duo (M2D), which outperformed previous techniques, achieving a weighted accuracy of 0.832 and an unweighted average recall of 0.713 [7]. The effectiveness of ensembling M2D with other models demonstrates the broader applicability of general-purpose audio representations in heart sound analysis.

The continuous evolution of AI-driven heart sound analysis has also seen the integration of traditional ML methods with deep learning. A study focused on time-frequency heat maps combined with a deep CNN to detect abnormalities in heart sounds achieved commendable performance, balancing sensitivity and specificity—an essential aspect of cost-sensitive medical diagnostics [11]. Furthermore, Cardi-Net, a deep learning model combining CNN and power spectrogram analysis, was introduced to extract discriminative features from PCG signals for the multi-class classification of four cardiac disorders without pre-processing or feature engineering [13]. Enhanced by data augmentation and 10-fold cross-validation, Cardi-Net achieved 98.88% accuracy, making it suitable for real-time use across various platforms, including cloud services and mobile apps.

In a different approach, the researchers transformed PCGs into spectral images that preserved the topological structure of the original data. This transformation allowed them to leverage the power of deep convolutional neural networks (CNNs) for feature extraction. To enhance the model's performance, data augmentation techniques were employed to increase the diversity of the training data. Additionally, transfer learning was utilized to fine-tune pre-trained CNN architectures, enabling the model to learn from existing knowledge [12]. To capture the temporal dynamics of cardiac murmurs, a recurrent neural network (RNN) was integrated into the architecture. This hybrid approach resulted in a significant improvement in accuracy, achieving a remarkable 94.01% in automatic cardiac murmur detection without the need for manual segmentation. The performance of an Adaptive Neuro-Fuzzy Inference System (ANFIS) was also evaluated for detecting abnormal cardiac valve sounds using spectral analysis features. After de-noising and feature extraction through High Order Spectral (HOS) analysis, the ANFIS model achieved classification accuracy between 63-89%, highlighting its potential in specific diagnostic contexts [14]. Another significant development is the creation of a portable, low-cost system for early detection of valvular heart abnormalities, such as arrhythmias and murmurs. Designed for use by untrained frontline health workers, this system processes stethoscope sounds into spectrograms for classification via cloud-based CNN models, achieving a 95% average classification accuracy [15]. Its validation with real-life heart sounds collected using a low-cost digital stethoscope demonstrates its promise as a comprehensive diagnostic tool for enhancing healthcare in developing regions.

Moreover, studies have explored the combination of PCG and ECG waveforms for enhanced disease screening through a novel dual-convolutional neural network approach. This method introduces both record-wise and sample-wise evaluation frameworks, showing that integrating ECG and PCG data significantly outperforms single-modality methods, leveraging transferable features from separately collected ECG and PCG waveforms for improved classification accuracy [16]. The application of ensemble models combining random forest and extreme gradient boost for heart sound classification has also been explored, with the ensemble model using Moth Flame Optimization (MFO) further improving results, reaching 89.53% accuracy, 0.9 F1 score, and 0.95 AUC [17].

The field has also seen the introduction of AI-driven heart monitoring devices that screen and identify heart sounds, transmitting data to healthcare providers through the Internet of Things (IoT) [18]. These systems, which employ LSTM architectures, enable patients to self-monitor their heart health, offering a novel approach to managing coronary conditions. On the other hand, the challenge of detecting heart disease from heart sound signals with imbalanced training and testing sets has also been addressed by developing ML models using features extracted from Discrete Wavelet Transform (DWT) and Mel-Frequency Cepstral Coefficients (MFCC). The study explored various models, including Random Forest and Extreme Gradient Boost, achieving high accuracy and AUC, particularly when using an ensemble model with Moth Flame Optimization [17]. Finally, a combination of conventional feature engineering and deep learning has been employed to classify normal and abnormal heart sounds automatically. Initially, 497 features were extracted from eight domains, which were then fed into a CNN. To prevent overfitting, fully connected layers were replaced with a global average pooling layer, and class weights were adjusted to address class imbalance [19]. Using stratified five-fold cross-validation, the method achieved a mean accuracy of 86.8%, sensitivity of 87%, specificity of 86.6%, and a Matthews correlation coefficient of 72.1%, striking a balance between sensitivity and specificity.

III. METHODOLOGY

The flowchart depicted in Fig. 1 illustrates the comprehensive methodology employed in this study. It details the entire process, from data acquisition through to the final model evaluation, highlighting the structured and methodical approach taken to develop, refine, and assess the performance of the heart murmur detection models.

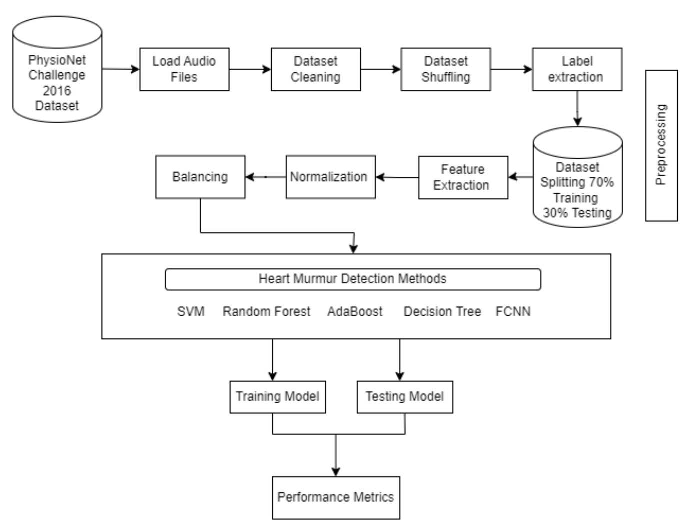

Fig. 1. Heart murmur detection model.

A. Data Collection

The PhysioNet Challenge 2016 dataset was utilized for this study [20], comprising heartbeat sounds collected from a diverse range of patients in clinical settings using electronic stethoscopes. The recordings, saved in .wav format with a sampling rate of 2 kHz, include normal heartbeats, murmurs, and other pathological conditions like extrasystole and gallop rhythms. These high-quality recordings were gathered from multiple clinical sites worldwide, ensuring a broad spectrum of acoustic environments and patient conditions. To maintain consistency and quality, standard recording protocols were adhered to, including consistent microphone placement and controlled ambient noise levels.

Each audio file is accompanied by metadata that includes labels indicating the presence or absence of heart murmurs, as diagnosed by expert cardiologists. The dataset is diverse, with recordings from patients across various age groups, genders, and clinical histories, which is crucial for developing models that generalize well across different populations. The dataset includes 3126 recordings from 764 patients, captured from different auscultation points. These recordings vary in length, allowing for the study of heart sounds over different time intervals. Despite challenges such as patient movement, breathing, and background noise, the dataset remains one of the most comprehensive publicly available collections of heart sounds, making it invaluable for research in heart murmur detection and related fields.

B. Data Preprocessing

The preprocessing of the dataset involved several critical steps to ensure the quality and suitability of the data for model training and evaluation. First, the audio files were loaded using the Librosa library in Python, which is specifically created for audio and music analysis. It offers essential tools for handling audio data, simplifying the extraction and manipulation of features, signal analysis, and the execution of tasks such as beat tracking, pitch detection, and sound classification. During this stage, any corrupted or incomplete files were identified and discarded to maintain the integrity of the dataset.

To eliminate any potential ordering bias and ensure the data is randomized, the entire dataset was shuffled. This process is essential to prevent the model from picking up unintended patterns based on the order of the data, which could skew the learning process. Shuffling helps the model avoid overfitting by not relying on the sequence of data, thereby enhancing its ability to generalize to new, unseen examples and improving overall model performance.

Following the shuffling process, labels indicating the presence or absence of heart murmurs were extracted from the accompanying metadata to create the target variable for classification tasks. The labeling method involved using predefined criteria from the metadata files, which typically included annotations from medical professionals or automated detection algorithms. This extraction was performed with meticulous care to ensure that each label was correctly aligned with its corresponding audio recording. Accurate alignment of labels and recordings is essential for reliable model training and evaluation, as it ensures that the model is learning from properly matched data and enhances the quality of the classification process.

Finally, the dataset was subsequently divided into training and test sets using stratified sampling, with 70% of the data allocated to the training set and 30% to the test set. Stratified sampling was employed to ensure that both subsets accurately reflected the original distribution of positive and negative cases. This method is crucial for addressing the dataset's inherent imbalance, as it prevents the creation of subsets that could disproportionately favor one class over the other. By maintaining proportional representation in both the training and test sets, stratified sampling helps the model learn from a balanced perspective, which enhances its ability to generalize effectively to new, unseen data. Furthermore, this approach supports a more reliable evaluation of the model's performance, reducing the risk of biased results and ensuring that both positive and negative cases are adequately represented in the testing phase.

C. Features Extraction

As discussed earlier, MFCC was selected for feature extraction due to its proven ability to represent audio signals in a form that is highly compatible with ML and DL models. MFCCs effectively capture the timbral characteristics of audio, which are crucial for distinguishing subtle variations, such as heart murmurs, that may be indicative of underlying health conditions. In addition to MFCCs, other features were extracted to provide a comprehensive representation of the audio signals. Chroma features, which capture the harmonic content of the audio, were computed, and their mean and standard deviation were calculated. Spectral contrast features, representing the difference in amplitude between peaks and valleys in the sound spectrum, were also extracted along with their mean and standard deviation. Lastly, Tonnetz (tonal centroid features), which capture the tonal properties of the audio, were computed, and their mean and standard deviation were included.

During this process, each audio recording was segmented into short frames of 25ms, with a 10ms overlap to ensure that transient features were not missed. For each frame, 13 MFCCs were computed using the librosa library, based on a filter bank that mimics the human ear's perception of sound. The resulting coefficients were chosen because they offer a balance between computational efficiency and the richness of information captured. After extracting the MFCCs, the mean and standard deviation across all frames were calculated, resulting in a fixed-length feature vector for each recording. This approach ensures that the variability within each recording is captured, while also reducing the dimensionality of the data, making it more manageable for ML and DL models.

In total, 130 features were extracted from each audio file, comprising 40 MFCCs (mean and standard deviation), 24 chroma features (mean and standard deviation), 14 spectral contrast features (mean and standard deviation), and 12 Tonnetz features (mean and standard deviation). These features collectively provide a rich and detailed representation of the heart sound recordings, enabling effective analysis and classification by the ML and DL models. MFCCs are widely favored in audio processing because they provide a compact representation of the power spectrum of sound, making them ideal for detecting subtle anomalies like heart murmurs.

Additionally, the choice of 13 coefficients aligns with common practices in speech and audio processing, where it has been empirically shown that this number provides sufficient detail for accurate modeling while avoiding overfitting. The extracted feature vectors were then normalized and fed into the ML and DL models. This step ensures consistency across recordings and enhances the models' ability to generalize from the training data to unseen examples. By leveraging these diverse features, the models can better capture the nuanced differences between normal heart sounds and those that indicate murmurs, ultimately improving the accuracy of the detection system.

D. Normalization and Balancing

Following the feature extraction process, the audio signals were normalized to achieve a zero mean and unit variance. This normalization step is crucial for ensuring consistency across the input data by removing scale differences between features. Such standardization not only enhances the overall performance of the model but also facilitates faster and more stable convergence during the training phase. By normalizing the data, we mitigate potential biases that could arise from varying signal amplitudes, thereby improving the reliability and accuracy of the model's predictions.

Moreover, an analysis of the dataset revealed a significant imbalance between normal and abnormal recordings, with normal recordings being more prevalent. This disparity could lead to the machine learning models favoring the majority class, resulting in suboptimal performance on the minority class. To address this issue, we applied techniques like the Synthetic Minority Over-sampling Technique (SMOTE) and Random Over-Sampling. SMOTE works by generating synthetic samples for the minority class. It does this by interpolating between existing minority class samples, creating new data points that lie along the line segments connecting nearest neighbors in the feature space. This method not only increases the number of minority class samples but also introduces more variability, helping to reduce the risk of overfitting. On the other hand, Random Over-Sampling (ROS) involves duplicating existing minority class samples to balance the dataset. While ROS is straightforward and effective, it can sometimes lead to overfitting because it doesn't introduce new information into the model. By combining SMOTE and ROS, we ensured a more balanced dataset, which is crucial for training models that can make unbiased and reliable predictions. This balanced approach helps the model to better generalize across both the majority and minority classes, ultimately leading to improved detection of abnormal cases in the dataset.

E. Applying the Used Methods

1) ML and DL Algorithms: The final step in the methodology involves applying a range of machine learning (ML) and deep learning (DL) techniques to the preprocessed dataset to detect heart murmurs. Specifically, we employed Support Vector Machine (SVM) [21], Random Forest (RF) [22], AdaBoost [23], Decision Tree [24], and a Fully Convolutional Neural Network (FCNN) [25] for this classification task. Each of these algorithms brings unique strengths: SVM is effective in handling high-dimensional spaces, RF and AdaBoost are robust against overfitting, and FCNN excels in learning complex patterns directly from the raw audio data. To rigorously evaluate the performance of each model in detecting heart murmurs, we calculated key metrics including accuracy, precision, recall, and F1 score. These metrics offer a comprehensive evaluation, ensuring not only the overall correctness of the models but also their capability to balance the trade-offs between false positives and false negatives, an essential consideration in the accurate detection of heart murmurs.

2) Hyperparameter tuning and architectural design: Hyperparameter tuning is a crucial aspect of any machine learning (ML) and deep learning (DL) application, as it directly impacts the model's performance by finding the optimal settings that allow the algorithm to best capture the underlying patterns in the data. For instance, in the case of Support Vector Machine (SVM), hyperparameter tuning was meticulously performed using GridSearchCV to optimize critical parameters such as C (the regularization parameter) and gamma (the kernel coefficient). The parameter C controls the trade-off between achieving a low error on the training data and minimizing the model's complexity, while gamma defines the influence of a single training example. Exploring a range of values, including C: [0.01, 0.1, 1, 10] and gamma: ['scale', 'auto'], ensured a comprehensive search of the parameter space, thereby enhancing the model's ability to generalize to new data. This systematic exploration is essential because it allows the model to adapt to the specific characteristics of the dataset, leading to improved accuracy and robustness.

Similarly, the decision tree classifier was finetuned using GridSearchCV as well to optimize parameters such as max_depth, min_samples_split, and min_ _samples_leaf. These parameters are critical for controlling the tree's growth and complexity—max_depth limits the depth of the tree to prevent overfitting, while min_samples_split and min_samples_leaf dictate the minimum number of samples required to split an internal node and to be at a leaf node, respectively. By testing values for max_depth: [5, 10, 15], min_samples_split: [2, 5, 10], and min_samples_leaf: [1, 2, 5], the decision tree was carefully tailored to achieve the optimal balance between model complexity and predictive power, ensuring that the tree was neither too simplistic nor overly complex.

In the deep learning model, constructed using TensorFlow/Keras, the architecture was designed with careful consideration of both complexity and computational efficiency. The Sequential model began with a dense input layer consisting of 256 neurons, followed by three hidden layers with 64, 32, and 16 neurons, each utilizing ReLU activation functions. To mitigate the risk of overfitting, a dropout rate of 0.5 was introduced between layers, which helped in maintaining the model's ability to generalize by randomly disabling a fraction of the neurons during training. The output layer, a dense layer with a single neuron and a sigmoid activation function, was designed to output the probability of heart murmur presence. The architectural choices were driven by the need to balance the model's ability to learn intricate patterns within the data while avoiding excessive complexity that could lead to overfitting.

The model was compiled using the Adam optimizer and binary cross-entropy loss function, chosen for their efficiency and effectiveness in binary classification tasks. Training was conducted over 100 epochs with a batch size of 32, ensuring sufficient learning while maintaining computational feasibility. Regularization techniques, such as early stopping (with patience set to 10) and learning rate reduction (with a factor of 0.5 and patience set to 5), were employed to prevent overfitting and ensure that the model converged to an optimal solution. These strategies collectively contributed to building a robust and efficient model capable of accurately detecting heart murmurs from the given dataset.

IV. Results and Discussion

The results of this study underscore the significant advancements achieved through both traditional ML models and a DL model in detecting heart murmurs from audio recordings. The comparative analysis, as presented in the provided Table I, reveals distinct differences in performance between the ML models—Support Vector Machine (SVM), Random Forest, AdaBoost, and Decision Tree—and the deep learning model, specifically the FCNN.

The SVM model achieved an accuracy of 0.85, with corresponding F1-score, precision, and recall values of 0.84, 0.84, and 0.85, respectively as depicted in Fig. 2. While these results provide a solid baseline, they indicate that SVM struggled to handle the complexity of the data, particularly in terms of balancing precision and recall. This suggests that the model's reliance on a hyperplane for classification may not be the most effective strategy for high-dimensional, complex heart sound data, which exhibits non-linear relationships that require more flexible learning methods.

Similarly, the Decision Tree model, with an accuracy of 0.89 and matching precision and recall values, shown in Fig. 3, performed better than SVM but still exhibited limitations in fully capturing the intricate patterns within the data. This is expected, as single-tree models are prone to overfitting and fail to generalize well, especially when dealing with highly variable heart sound signals.

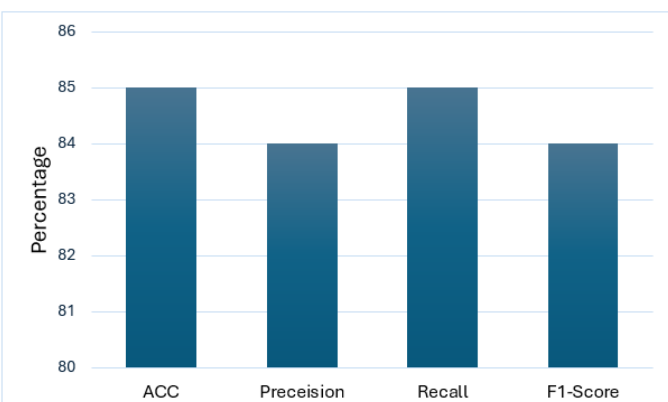

Fig. 2. SVM classification results.

In contrast, the ensemble methods, Random Forest and AdaBoost, demonstrated enhanced performance, particularly in terms of their robustness against over-fitting. The Random Forest model achieved an accuracy of 0.87, with an F1-score of 0.88, precision of 0.90, and recall of 0.87 as illustrated

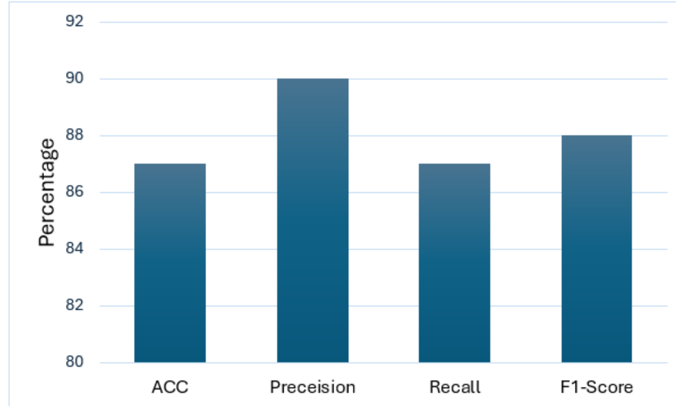

Fig. 3. Random forest classification results.

in Fig. 4. These results highlight the model's ability to generalize well across the dataset, benefiting from the ensemble approach's capacity to combine multiple decision trees and reduce variance. AdaBoost, with an accuracy of 0.88 and consistent F1-score, precision, and recall of 0.88, highlighted in Fig. 5, illustrates the potential of boosting techniques in improving model performance by focusing on misclassified instances. The performance improvement of AdaBoost suggests that iterative reweighting of data points can effectively guide the learning process toward difficult-to-classify cases, making it particularly valuable for datasets with subtle variations, such as heart murmurs.

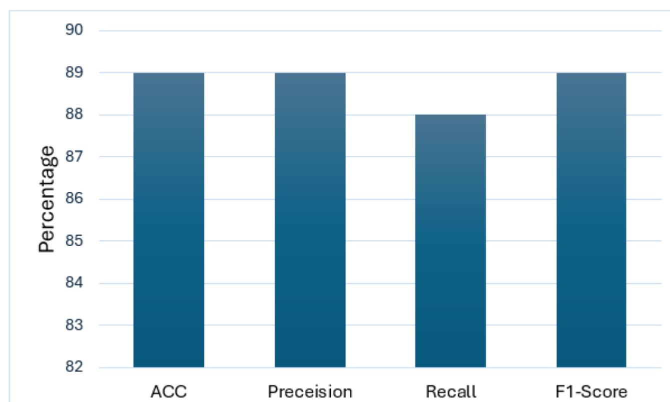

Fig. 4. AdaBoost classification results.

However, the most significant improvement was observed with the FCNN deep learning model, which outperformed all traditional ML models by a considerable margin. The FCNN achieved an impressive accuracy of 0.99, with an F1-score of 0.94, precision of 0.94, and recall of 0.96. These results displayed in Fig. 6 demonstrate the deep learning model's superior capability in capturing the complex and subtle features within the heart sound recordings that the traditional models struggled to detect. One of the key advantages of the FCNN is its ability to automatically extract hierarchical features from the raw audio data, which is particularly valuable in this study for identifying minute acoustic variations in heart murmurs that might be overlooked by traditional feature extraction methods.

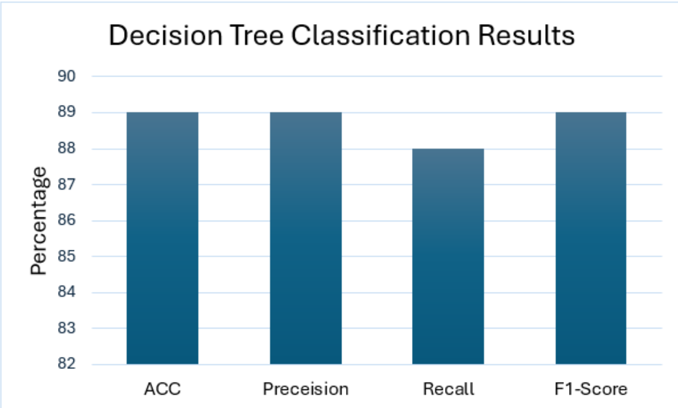

Fig. 5. Decision tree classification results.

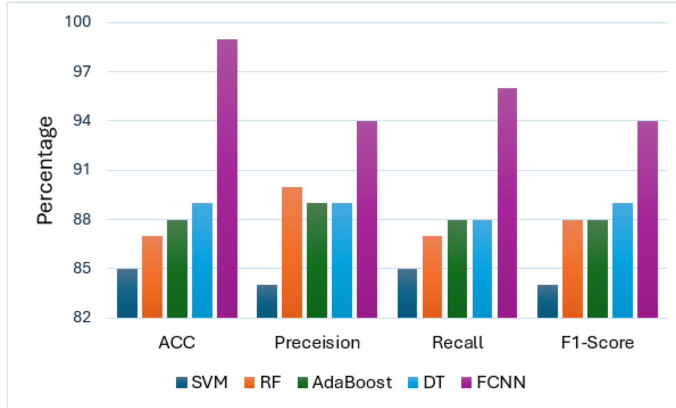

Fig. 7. Classification results of used algorithms.

The architecture of the FCNN, with its multiple hidden layers, ReLU activation functions, and dropout mechanisms, allowed for an effective learning process that minimized overfitting and maximized predictive accuracy. The use of regularization techniques such as early stopping and learning rate reduction further optimized the model's performance, ensuring that it remained both accurate and generalizable. This suggests that deep learning models not only outperform traditional approaches but also maintain stability across diverse datasets, a crucial factor for clinical adoption.

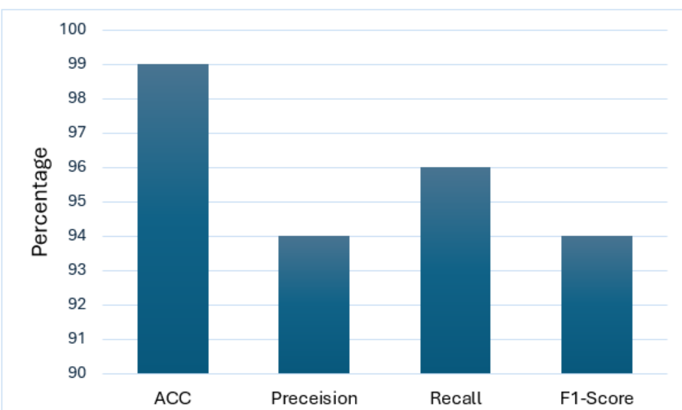

Fig. 6. FCNN Classification results.

Overall, the classification results are comprehensively displayed in Fig. 7, effectively highlighting the performance differences across all models. The drastic improvement seen in FCNN emphasizes that deep learning architectures, with their ability to handle complex feature interactions, could revolutionize heart murmur diagnostics, potentially outperforming traditional auscultation and even current ML-based methods.

These results, as compared in Table I underscore the transformative potential of deep learning models in the evolving landscape of heart murmur detection, signaling a paradigm shift from traditional machine learning approaches. The traditional models, such as Support Vector Machines (SVM), Decision Trees, and even ensemble methods like Random Forest and AdaBoost, provided a valuable baseline in this study. They demonstrated solid performance in terms of accuracy, precision, recall, and F1-score, especially when paired with techniques to manage data complexities and imbalances. However, despite these efforts and improvements, these models still fell short of the performance levels achieved by the Fully Convolutional Neural Network (FCNN).

TABLE I. Performance Comparison of ML and DL Models for Heart Murmur Detection

Model used	Accuracy	F1-score	Precision	Recall
		ML		
SVM	0.85	0.84	0.84	0.85
Random Forest	0.87	0.88	0.90	0.87
AdaBoost	0.88	0.88	0.89	0.88
Decision Tree	0.89	0.89	0.89	0.89
		DL		
FCNN	0.99	0.94	0.94	0.96

The superior performance of the FCNN suggests that deep learning models have a distinct advantage in handling the complexities inherent in medical data, particularly in tasks like heart murmur detection. Unlike traditional models, which often rely on manual feature extraction and struggle with high-dimensional data, deep learning models are capable of automatically learning intricate patterns from raw data. This ability is especially crucial in medical diagnostics, where subtle variations in data, such as the nuanced differences in heart sound recordings, can significantly impact patient outcomes. The FCNN's architecture, with its deep layers, ReLU activations, and regularization techniques, enabled it to capture these subtleties effectively, leading to higher accuracy and more reliable predictions.

Moreover, the FCNN's capacity to process vast amounts of data with minimal need for extensive feature engineering underscores a significant advantage of deep learning in the realm of medical diagnostics. As the healthcare industry continues to produce enormous volumes of data—from electronic health records and imaging to sensor-based monitoring—models that can efficiently analyze and learn from this data will become increasingly indispensable. The implications of these capabilities are substantial. As deep learning models consistently demonstrate their effectiveness in handling complex diagnostic tasks, they are poised to become fundamental components of medical practice, significantly improving the precision and speed of diagnoses.

In the specific context of heart murmur detection, this technological advancement could lead to the earlier and more accurate identification of potentially life-threatening cardiovascular conditions, thereby improving patient outcomes and streamlining healthcare delivery. Furthermore, the inherent scalability of deep learning models makes them particularly well-suited for integration into comprehensive healthcare systems, including telemedicine platforms and automated diagnostic tools. This scalability ensures that their benefits can be extended across diverse healthcare settings, from remote clinics to large urban hospitals, further amplifying their impact on patient care and the overall efficiency of medical services.

V. Conclusion

This comparative study, which included both machine learning (ML) and deep learning (DL) methods, highlights the significant advancements and potential of using these techniques for the early diagnosis and detection of heart murmurs. The study utilized the PhysioNet Challenge 2016 dataset, a comprehensive collection of heartbeat sounds gathered from a diverse range of patients using electronic stethoscopes in clinical settings. This dataset, comprising over 3,000 recordings from 764 patients, provided a robust foundation for training and evaluating various models. The methodology involved several key steps, starting with the preprocessing of audio data using the Librosa library, where features such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, spectral contrast, and Tonnetz were extracted to capture the essential characteristics of the heart sounds. Following feature extraction, the dataset was divided into training and test sets using stratified sampling to ensure balanced representation of positive and negative cases. The study then applied a range of ML and DL algorithms, including Support Vector Machine (SVM), Random Forest, AdaBoost, Decision Tree, and a Fully Convolutional Neural Network (FCNN), to classify heart murmurs. Hyperparameter tuning was meticulously performed using GridSearchCV to optimize model performance, ensuring that each algorithm was tailored to the specific characteristics of the dataset. The FCNN, in particular, demonstrated a substantial improvement in accuracy and reliability over traditional methods, underscoring the potential of DL models in this domain.

The results suggest that integrating AI-powered diagnostic tools into clinical practice could lead to earlier and more precise diagnoses, thereby improving patient outcomes and reducing the global burden of cardiovascular diseases. However, several limitations must be acknowledged. The primary challenge lies in the reliance on the dataset, which, while comprehensive, may not fully represent the variability encountered in real-world clinical settings, where factors such as diverse demographic characteristics, comorbidities, and recording environments could affect the model's generalizability. Additionally, despite the superior performance of the deep learning model, its "black box" nature poses challenges for clinical adoption, as it makes it difficult for clinicians to interpret the rationale behind specific predictions, potentially hindering trust in the model.

Furthermore, the study focused primarily on heart sounds recorded under controlled conditions, without thoroughly addressing the impact of real-world noise and artifacts, which could degrade the model's performance in actual clinical environments. To overcome these limitations and advance the field further, future research should focus on acquiring more diverse and representative datasets that include a broader range of patient demographics and clinical conditions, thus enhancing the model's generalizability. Additionally, developing personalized heart murmur detection models that take into account individual patient characteristics, such as medical history and genetic data, could lead to even more accurate and relevant predictions. Finally, efforts should be made to enhance the interpretability of deep learning models through explainable AI techniques, which could provide clinicians with better insights into the model's decision-making process and facilitate greater integration into clinical practice.

Acknowledgment

The authors extend their appreciation to the Arab Open University and AlYamamah University for funding this work.

References

[1] World Health Organization, "Cardiovascular Diseases (CVDs)," World Health Organization, 2021, accessed: 2024-08-21. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] S. S. Martin et al., "2024 heart disease and stroke statistics: a report of US and global data from the American Heart Association," Circulation, vol. 149, no. 8, pp. e347–e913, 2024.

[3] E. Mejia and S. Dhuper, Innocent Murmur, updated 2023 sep 4 ed., ser. StatPearls. Treasure Island (FL): StatPearls Publishing, 2024. Available from: https://www.ncbi.nlm.nih.gov/books/NBK507849/.

[4] S. Das and S. Dandapat, "Heart murmur severity stages classification using multi-kernel residual cnn," IEEE Sensors Journal, 2024.

[5] M. F. A. B. Hamza and N. N. A. Sjarif, "A comprehensive overview of heart sound analysis using machine learning methods," IEEE Access, 2024.

[6] M. Jobayer et al., "FunnelNet: an end-to-end deep learning framework to monitor digital heart murmur in real-time," arXiv preprint arXiv:2405.09570, 2024.

[7] D. Niizumi et al., "Exploring pre-trained general-purpose audio representations for heart murmur detection," arXiv preprint arXiv:2404.17107, 2024.

[8] T. Wu et al., "Heart murmur quality detection using deep neural networks with attention mechanism," Applied Sciences, vol. 14, no. 15, p. 6825, 2024.

[9] Y.-T. Tsai et al., "Heart murmur classification using a capsule neural network," Bioengineering, vol. 10, no. 11, p. 1237, 2023.

[10] S. B. Shuvo et al., "Cardioxnet: A novel lightweight deep learning framework for cardiovascular disease classification using heart sound recordings," IEEE Access, vol. 9, pp. 36,955–36,967, 2021.

[11] J. Rubin et al., "Recognizing abnormal heart sounds using deep learning," arXiv preprint arXiv:1707.04642, 2017.

[12] Chorba, John S., Shapiro, Avi M., Le, Le, Maidens, John, Prince, John, Pham, Steve, Kanzawa, Mia M., Barbosa, Daniel N., Currie, Caroline, Brooks, Catherine, et al. "Deep learning algorithm for automated cardiac murmur detection via a digital stethoscope platform." *Journal of the American Heart Association*, vol. 10, no. 9, 2021, e019905. Am Heart Assoc.

[13] J. S. Khan et al., "Cardi-net: A deep neural network for classification of cardiac disease using phonocardiogram signal," Computer Methods and Programs in Biomedicine, vol. 219, p. 106727, 2022.

[14] B. Al-Naami et al., "A framework classification of heart sound signals in physionet challenge 2016 using high order statistics and adaptive neuro-fuzzy inference system," IEEE Access, vol. 8, pp. 224852–224859, 2020.

[15] K. K. Singh and S. S. Singh, "An artificial intelligence based mobile solution for early detection of valvular heart diseases," in 2019 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), IEEE, 2019, pp. 1–5.

[16] R. Hettiarachchi et al., "A novel transfer learning-based approach for screening pre-existing heart diseases using synchronized ecg signals and heart sounds," in 2021 IEEE international symposium on circuits and systems (ISCAS), IEEE, 2021, pp. 1–5.

[17] A. Rath et al., "Development and assessment of machine learning based heart disease detection using imbalanced heart sound signal," Biomedical Signal Processing and Control, vol. 76, p. 103730, 2022.

[18] U. Dampage et al., "AI-based heart monitoring system," in 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), IEEE, 2021, pp. 1–6.

[19] F. Li et al., "Classification of heart sounds using convolutional neural network," Applied Sciences, vol. 10, no. 11, p. 3956, 2020.

[20] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," Circulation, vol. 101, no. 23, pp. e215–e220, 2000 (June 13).

[21] D. A. Pisner and D. M. Schnyer, "Support vector machine," in Machine Learning, Elsevier, 2020, pp. 101–121.

[22] G. Biau and E. Scornet, "A random forest guided tour," Test, vol. 25, pp. 197–227, 2016.

[23] R. E. Schapire, "Explaining adaboost," in Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik, Springer, 2013, pp. 37–52.

[24] Y.-Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction," Shanghai archives of psychiatry, vol. 27, no. 2, pp. 130, 2015.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.